# Improving Command Selection In Smart Environments By Exploiting Spatial Constancy

A Thesis Submitted to the College of

Graduate Studies and Research

In Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy

In the Department of Computer Science

University of Saskatchewan

Saskatoon

By

**Adrian Reetz**

## Permission to Use

In presenting this dissertation in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis/dissertation work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis/dissertation.

## Disclaimer

Reference in this thesis/dissertation to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis/dissertation in whole or part should be addressed to:

*Head of the Department of Computer Science*
*University of Saskatchewan*
*110 Science Place*
*Saskatoon, Saskatchewan S7N 5C9*
*Canada*

or

*Dean*
*College of Graduate Studies and Research*
*University of Saskatchewan*
*107 Administration Place*
*Saskatoon, Saskatchewan S7N 5A2*
*Canada*

## Abstract

With the a steadily increasing number of digital devices, our environments are becoming increasingly smarter: we can now use our tablets to control our TV, access our recipe database while cooking, and remotely turn lights on and off. Currently, this Human-Environment Interaction (HEI) is limited to in-place interfaces, where people have to walk up to a mounted set of switches and buttons, and navigation-based interaction, where people have to navigate on-screen menus, for example on a smart-phone, tablet, or TV screen. Unfortunately, there are numerous scenarios in which neither of these two interaction paradigms provide fast and convenient access to digital artifacts and system commands. People, for example, might not want to touch an interaction device because their hands are dirty from cooking: they want device-free interaction. Or people might not want to have to look at a screen because it would interrupt their current task: they want system-feedback-free interaction. Currently, there is no interaction paradigm for smart environments that allows people for these kinds of interactions.

In my dissertation, I introduce *Room-based Interaction* to solve this problem of HEI. With room-based interaction, people associate digital artifacts and system commands with real-world objects in the environment and point toward these real-world proxy objects for selecting the associated digital artifact. The design of room-based interaction is informed by a theoretical analysis of navigation- and pointing-based selection techniques, where I investigated the cognitive systems involved in executing a selection. An evaluation of room-based interaction in three user studies and a comparison with existing HEI techniques revealed that room-based interaction solves many shortcomings of existing HEI techniques: the use of real-world proxy objects makes it easy for people to learn the interaction technique and to perform accurate pointing gestures, and it allows for system-feedback-free interaction; the use of the environment as flat input space makes selections fast; the use of mid-air full-arm pointing gestures allows for device-free interaction and increases awareness of other's interactions with the environment.

Overall, I present an alternative selection paradigm for smart environments that is superior to existing techniques in many common HEI-scenarios. This new paradigm can make HEI more user-friendly, broaden the use cases of smart environments, and increase their acceptance for the average user.

## Acknowledgements

I would like to thank my supervisors Carl Gutwin for his unconditional support and for sacrificing so much of his time in order to guide me for the last seven years.

I would also like to thank the members of my committee for their interest, guidance, and encouragement. Thank you to Ralph Deters, Ian Stavness, Scott Bell (cognate), and to my external examiner Derek Reilly.

The work in this dissertation has benefited from many great collaborations. I would like to thank Regan Mandryk and Miguel Nacenta for providing me with fantastic guidance on several projects. I would also like to thank the members of the Interaction Lab for providing an amazing environment in which to work and learn.

I could not have made it through the last seven years without the encouragement and support of my friends. Thank you to Zoe Zhou, Malcolm Lucy, Andre Doucette, and Nelson Wong.

Finally, I want to thank my parents, Eva and Hartmut for their unconditional love and support throughout my entire life. Ich bedanke mich bei meinen Eltern Eva und Hartmut für Ihre bedingungslose Liebe un Unterstüzung.

## Dedication

I dedicate my dissertation to my parents: my mother, Eva Reetz, and my dad, Hartmut Reetz, who is not with us anymore but will forever live on in my heart and thoughts.

Ich widme meine Dissertation meinen Eltern: meiner Mutter Eva Reetz und meinem Vater Hartmut Reetz, der nicht mehr unter uns weilt aber für immer in meinem Herzen und meinen Gedanken weiterleben wird.

## Publications from this Dissertation

Adrian Reetz and Carl Gutwin. 2014. Making big gestures: effects of gesture size on observability and identification for co-located group awareness. In *Proceedings of the 32nd conference on Human Factors in Computing Systems – CHI '14*. ACM Press, New York, NY, USA, 4087–4096. http://doi.org/10.1145/2556288.2557219

## Permissions

Figure 3-1: Reprinted from Ishii and Ullmer, 1997, with permission from ACM.

Figure 3-2: Reprinted from reactable.com, with kind permission from Reactable Systems, SL.

Figure 4-1: Reprinted from Baudel and Beaudouin-Lafon, 1993, with permission from ACM.

Figure 4-2: Reprinted from Gustafson et al., 2010, with permission from ACM.

Figure 5-1: Reprinted from Wilson and Shafer, 2003, with permission from ACM.

Figure 5-2: Reprinted from Wilson and Pham, 2003, with kind permission from IOS Press.

Figure 6-1: Reprinted from Li et al., 2009, with permission from ACM.

Figure 6-2: Reprinted from Cockburn et al., 2011, with kind permission from Elsevier.

Figure 10-1: Reprinted from Hinckley et al., 1994, with permission from ACM.

Figure 10-2: Reprinted from Robertson et al., 1998, with permission from ACM.

## Table of Contents

## Table of Tables

## Table of Figures

# Chapter 1 Introduction

The number of interactive digital devices in people's environments is steadily increasing. Current technology allows people, for example, to stream music and videos into their living rooms; to share shopping lists between fridge, computer, and smart phone; and to control a variety of electronic devices, such as lights, blinds, and thermostats, remotely. This ubiquity of devices creates *smart environments*, and with it the need for methods that allow people to control all the *digital artifacts* in those environments. I call this **Human-Environment Interaction** (HEI).

*Artifacts* are "things" that users can control or access. Examples include simple electronic devices (e.g., ceiling lamps, heaters), appliances (e.g., stoves, microwaves), entertainment devices (e.g., TV sets, gaming consoles), and digital media (TV stations, movies, video games, podcasts, websites). An artifact is *digital* when it is part of a computer network and thus can be controlled remotely by the user. This can mean that "dumb" devices (e.g., radiators) have to be augmented with digital technology, or that unconnected devices (e.g., microwaves) have to be augmented with network interfaces.

I define *environments* as a confined physical space, such as a room (e.g., living room, kitchen, or bedroom) or office space (e.g., office or cubicle). I focus my research on stationary environments that users spend an extended amount of time in and are thus familiar with (i.e., users have a good understanding about the spatial layout of the environment, the real-world objects within it, and the digital artifacts available). An environment is *smart* when it contains digital artifacts (i.e., when users can control artifacts in it remotely).

Human-Environment Interaction (HEI) allows people to achieve goals in smart environments, for example, turning on the lights to read a book, consulting an online recipe while cooking, or selecting a TV show for entertainment. It defines a subarea of Human-Computer Interaction (HCI) and particularly focuses on selection techniques for digital artifacts in smart environments. In the dissertation, I specifically limit HEI to artifact selection, which is choosing a single artifact from a larger group. Currently, people usually perform HEI through a control device. These devices can range from simple wall-mounted buttons and dials for controlling ceiling lights or room heaters to complex dedicated remote controls and on-screen displays (e.g., a TV or

smartphone) for browsing through media libraries. Other digital artifacts, such as microwaves, stoves, and alarm clocks, are controlled directly by walking up to the device and using an on-device interface. These examples reveal two properties that underlie most of current HEI techniques. First, interfaces may follow an in-place paradigm, i.e. they are fixed to a specific location in the environment, for example, a device or a wall. This in-place nature of interfaces requires users to physically move to the location of the interface in the environment. Second, interfaces may use a navigation paradigm, that is, they require users to navigate through some menu-based user interface. These navigation-based interfaces force users to shift their visual and cognitive attention to the user interface.

## 1.1 The Problem with Existing Techniques for Human-Environment Interaction

Interacting with artifacts in current smart environments is often difficult because the two aforementioned properties, in-place interaction and navigation-based interaction, cause three inherent problems.

### 1.1.1 Performance

The first problem is that current HEI techniques are either slow, inconvenient and disruptive, or require too much physical and cognitive effort from their users.

Consider a typical activity in a domestic smart environment: looking up a stored cooking recipe while cooking. With current HEI techniques, this activity might require people to stop their current cooking task, wash and dry their hands, and wake and unlock their tablet. This example demonstrates how current HEI techniques can slow down people: while the act of checking a recipe only requires a single glance and can happen within a few seconds, the surrounding HEI significantly prolongs this activity. The example also shows how current HEI techniques disrupt and inconvenience people: they have to perform procedures, such as cleaning their hands, for the sole purpose of interacting with the smart environment. In conclusion, current HEI techniques can make interaction with smart environments disruptive, inconvenient, and slow.

Another typical activity in modern living rooms is resuming a TV show that was watched earlier. With current navigation-based HEI techniques, people might have to locate and pick up a control device (e.g., smart phone or remote control), navigate through on-screen menus, find the correct

show, and select “resume”. As before, this example illustrates how current HEI techniques prolong the execution of an otherwise short and simple command, “resume TV show”. Conceptualizing and even verbalizing this command requires only a few seconds, but the act of finding the control device and navigating its menus prolongs the execution of the command gravely. This example also illustrates the cognitive effort required by current HEI techniques for giving a command to the smart environment: navigating menus demands people’s undivided visual attention and requires people to perform a complex cognitive and motoric activity. In conclusion, the inherent problems of current navigation-based HEI techniques can make interaction with smart environments slow and cognitively and physically costly.

The final example for problems with current HEI techniques is the simple and frequently performed activity of turning on the ceiling lights while reading a book. With current in-place HEI techniques, this activity might require people to put the book aside, walk over to a wall panel, find and flick the correct switch, return to their seat, and open the book again. This example shows how the in-place nature of this HEI technique requires users to move to the interface and then back to their original location, which takes time and makes HEI slow. While flicking the switch requires only a split second, with current in-place HEI techniques this process takes significantly more time to complete. In addition, this example demonstrates how in-place interfaces inherently require people’s physical effort for operation. Last, it exemplifies the disruptive property of in-place HEI techniques as people can only use them after completely suspending their current activity.

### 1.1.2 No Device-free Interaction

The second problem is that current HEI techniques do not provide device-free interaction, i.e. interaction in which people neither have to hold nor touch a control device. In all three previous scenarios, as well as many other use-cases, HEI plays a supporting role to a non-digital primary task: turning on the lights (HEI) while reading a book (primary), checking a recipe (HEI) while cooking (primary), confirming the outside temperature (HEI) while getting dressed (primary), or pausing the TV (HEI) while ironing (primary). In these scenarios, people are using their hands for completing the primary task. Current HEI techniques do not provide device-free interaction as they require users to hold or touch a HEI control device. This lack of device-free interaction makes current HEI techniques by design disruptive of people’s primary tasks and slow to

execute. Although similar to the first problem (disruptiveness), lack of device-free interaction deserves to be addressed explicitly as many situations in people's daily life prevent them from touching a control device, e.g., dirty hands (workshop), unhygienic working conditions (kitchen), or physical barrier (wearing gloves). Beyond the realm of domestic smart environments, there are other smart environments in which touch for device interaction is not feasible (e.g., sterile operating theaters or laboratories).

### 1.1.3 Visibility of Interaction

The final problem is that navigation-based HEI techniques hide interactions with the environments from other people in the group. HEI oftentimes happens in shared spaces, such as living rooms, where all present people can be affected by it. Thus, all present people have an interest in knowing when someone else is interacting with the smart environment. With in-place interfaces, the act of people moving to the interface is visible to all bystanders, and thus implicitly generates awareness. With navigation-based control devices, such as smart phones and tablets, in contrast, interaction is not visible to other people in the environment. This means that publicly important activities, such as changing the room temperature or setting an alarm clock, have to be explicitly verbally communicated to others because navigation-based device interaction does not create awareness for co-located people.

## 1.2 Room-based Interaction as Solution

I propose *Room-based Interaction* a new interaction paradigm that uses real-world objects in a smart environment as shortcuts (selection proxies) for digital artifacts and pointing gestures toward these objects in a single mid-air full-arm pointing gesture for digital artifact selection. An example would be that pointing at the living room window (real-world object) shows the local weather forecast (digital artifact) on the TV screen.

This novel interaction paradigm solves the three above-mentioned inherent problems with today's interaction with smart environments: lack of performance, lack of device-free interaction, and lack of public visibility.

Room-based interaction does not demand as much physical effort as in-place interfaces since it does not require people to move to the location of the interface, and it does not demand as much cognitive effort as navigation-based interfaces since it does not require people to navigate

complex menu hierarchies. For these reasons, room-based interaction also allows for faster and more convenient interaction than in-place and navigation-based interfaces. Because room-based interaction consists of a single mid-air full-arm pointing gesture, it additionally makes interaction less disruptive than in-place or navigation-based interfaces.

People can use room-based interaction device-free because, in contrast to in-place or navigation-based interfaces, people neither have to hold nor touch any device. In addition to device-free interaction, with sufficient training people might be able to use room-based interaction system-feedback- and eyes-free. System-feedback-free means that people can interact with the system without having to receive feedback from the system or having to look at system output during interaction. Navigation-based interfaces, for example, are generally not system-feedback-free as users have to look at a screen for navigating the menu structure. Eyes-free interaction means that people can interact with the system without having to direct visual attention away from their primary task. With room-based interaction, people might be able to point at the real-world proxy object relying entirely on body-intrinsic feedback and existing spatial knowledge.

Last, room-based interaction allows for publicly visible interactions with smart environments because its large mid-air full-arm pointing gestures are more visible to other people in the environment than the comparatively smaller gestures on navigation-based interfaces. Room-based interaction translates Don Norman's idea of "big gestures" into the realm of Human-Environment Interaction.

Room-based interaction combines two ideas: the use of real-world objects as selection proxies and the use of pointing gestures as selection mechanism. In detail, room-based interaction has the following capabilities and characteristics:

- Room-based interaction is memory-based, in particular based on relational and spatial memory. People use relational memory for remembering the association between digital artifacts and real-world proxy objects and spatial memory for finding the location of real-world proxies in the environment. Focusing on a memory-based interaction paradigm sets room-based interaction apart from in-place and navigation-based interaction techniques.
- With room-based interaction, associations between a digital artifact and real-world selection proxy can either be based on people's pre-existing meaningful semantic

knowledge, i.e. people can use a common and "natural" association, or they can create an entirely new meaningful abstract mapping between digital artifact and proxy object.

- Room-based interaction uses a single-level storage space, i.e. all real-world selection proxies are directly accessible and do not require browsing or navigating hierarchical storage structures. This is different from navigation-based interaction where people have to navigate hierarchical multi-level menus.
- Room-based interaction uses mid-air full-arm pointing gestures toward real-world selection proxies for device interaction. People use the same type of pointing gestures in face-to-face communication with others (deictic pointing gestures, for example, pointing at a car while saying "This car over there."). People already have the procedural knowledge for creating this type of gestures as it is a cornerstone of non-verbal human-human communication.
- With room-based interaction, people can choose to perform eyes-free pointing gestures, i.e. solely rely on spatial memory and proprioception for guiding their gestures toward selection proxies, or additionally acquire their pointing targets visually for increased pointing accuracy.

## 1.3 Scoping of this Dissertation

In this dissertation, I investigate the use of room-based interaction for digital artifact selection in smart environments with an emphasis on selection speed, device- and system-feedback-free interaction, and public visibility of interactions. While digital artifacts in the context of room-based interaction can be all types of device properties and digital resources that can be changed, accessed, or selected remotely, I will focus on discrete single command selections in my dissertation. In this type of interaction, people select a single digital artifact from a larger group of artifacts. I intentionally exclude selection of continuous values or sequences of selections. For example, I would consider selections such as "scroll recipe down by one screen" or "turn volume up", but not selections such as "set volume to X out of 100" or "type name of TV show to watch". The reason for this simplification is that this work is an initial exploration of the principles that underlie room-based interaction, most notably the combination of real-world proxy objects and full-arm pointing gestures. Using more complex selections would have added

another interaction layer on top of room-based interaction and thus obscured any investigation of the aforementioned basic principles.

Environments are physical spaces constrained by walls, a floor, and a ceiling; they are of an area and height that is commonly found in people's homes or typical office buildings. This means that I focus on environments of up to $10 \times 10\ m^2$ in area and up to $3\ m$ in height. An environment is smart when it contains digital artifacts. With selection, I refer to picking a single artifact out of a larger group. The selection tasks I am interesting in are typical tasks for domestic or office environment in that they are not time-critical and allow for mistakes to happen. These tasks include daily chores, such as cooking or ironing; leisure activities, such as reading, watching TV, or playing games; and productive activities, such as pair programming and repository commits. Publicly visible means visible to the group of people present in an aforementioned domestic environment. Groups can include up to around a dozen individuals, and people in a group normally know each other.

My dissertation is rooted in Human-Computer Interaction and is an investigation of interaction techniques, particularly room-based interaction. While I use the domain of smart domestic environments to set the background for my dissertation, motivate my research, and describe use cases for room-based interaction, it is important to understand that this domain is not in the focus of this dissertation. As a result, my dissertation is more concerned with, for example, human performance measures and other metrics related to Human-Computer Interaction and less concerned with, for example, hardware and implementation details and other potential research topics in smart environments.

## 1.4 Outline of this Dissertation

Although room-based interaction appears to have many advantages, there is little known about its effectiveness compared to other techniques, such as touch-based interaction. The goal of my dissertation is closing this gap in the existing body of knowledge.

In Chapter 2, I sketch out the design space that has been covered by previous research in human-computer interaction and show which parts of this space has not been deeply investigated. I also present existing knowledge from other fields that formed the foundation for my working hypotheses in the theory chapter (Chapter 3) and the user studies (Chapter 5 – Chapter 7).

In Chapter 3, I present a conceptual framework for assessing people's performance in pointing-based interaction. This framework is built on existing research presented in Chapter 2 and serves as a tool for conceptualizing the cognitive processes during the production of different types of pointing gestures and hypothesizing about people's performance. The working hypotheses for the three user studies (Chapter 5 – Chapter 7) are directly derived from this conceptual framework.

In Chapter 4, I describe different implementations of room-based interaction and demonstrate their feasibility with current hard- and software. With the theoretical framework and one concrete implementation in place (*Room Pointing*), I then report the results from three studies that determine the effectiveness of room-based interaction and compare it to existing techniques for interacting with smart environments.

In Chapter 5, I investigating the use of pointing as **selection mechanism** in room-based interaction in a user study. For this, I compare selection speed and accuracy of two touch-based interaction techniques, which can be considered today's default for interaction with smart environments, with two pointing-based interaction techniques. The two pointing-based techniques differ in that one uses traditional screen-based proxy objects (i.e., on-screen icons, similar to the Nintendo Wii), whereas the other, *Room Pointing*, an example for room-based interaction, uses real-world proxy objects. The main goal of this study is showing that people perform as good with pointing-based interaction techniques as with touch-based ones while former are providing the benefit of device-free interaction. Last, the first study informs the further investigations of mid-air full-arm pointing gestures.

In Chapter 6, I investigating the use of real-world objects as **selection proxies** in room-based interaction in a user study. Of particular interest here is quantifying how well people can remember associations between digital artifacts and real-world proxies. For this I compare *Room Pointing* with *Ray-casting Air-pointing*, a selection technique that uses virtual regions around a person as selection proxies (Cockburn, Quinn, Gutwin, Ramos, and Looser, 2011). Both interaction techniques use the same mid-air full-arm pointing gestures and differ solely in the type of selection proxy (real world objects versus virtual regions). This study also plays a crucial part in verifying the theoretical background of my work, which is founded on our current

understanding of learning spatial, semantic, and procedural information and our current knowledge about the influence of visual and proprioceptive feedback on human arm movement.

In Chapter 7, I investigate the communicative capabilities of mid-air full-arm pointing gestures in a user study. In particular, I am interested in the ability of these pointing gestures to create group and workspace awareness between co-located people. For this, gesture observation and identification are key components. I compare how well co-located people can observe and recognize phablet-sized touch-gestures, computer-screen-sized touch gestures, and mid-air full-arm pointing gestures.

In Chapter 8, I summarize the results from the three studies, evaluate how they match the hypotheses derived from the conceptual framework in Chapter 3, and tie them back into the larger context of room-based interaction. I close my dissertation in Chapter 9 with a short conclusion and an outlook of potential future work.

## 1.5 Contributions of my Dissertation

Overall, my dissertation makes four main contributions:

1. It **establishes the usefulness of real-world selection proxies for interacting with smart environments**. This idea was hinted at in previous research but has never been as thoroughly explored as in this work.
2. It shows how **mid-air full-arm pointing-based interaction facilitates device-free interaction as well as allows for eyes-free and (system-) feedback-free interaction**; it investigates the trade-off between selection accuracy, selection speed, and disruptiveness.
3. It **introduces a reference implementation for room-based interaction** called *Room Pointing* that exploits the idea of real-world selection proxies. By demonstrating the strengths and limitations of *Room Pointing* compared to other techniques, it **maps out the scope in which interaction designers can use room-based interaction techniques** to improve user's experience when interacting with smart environments.
4. It shows the general value of considering **gesture size as a factor for influencing the privacy of publicly performed gestures**. In particular, it demonstrates how mid-air full-arm pointing gestures can facilitate group and workspace awareness between co-located people.

# Chapter 2 Related Work

Room-based interaction describes selection techniques that are memory-based (i.e. use associations between digital artifacts and real-world proxy objects) and pointing-based (i.e. use mid-air full-arm pointing gestures toward real-world proxy objects for making selections). In this chapter, I give an overview of existing research about each of the components of room-based interaction: selection techniques, pointing gestures, proxy objects, and memory systems. First, I set the stage by describing smart environments, the application area of room-based interaction (2.1). I then present existing HCI literature on pointing-based selection techniques, gesture-based awareness creation, and the use of proxy objects (2.2). I then broaden the view by looking into psychology and kinesiology literature, where I first give a more rigorous definition of mid-air full-arm pointing gestures (2.3) and then lay out the cognitive processes involved in creating such gestures (2.4). Finally, I describe the different types of human memory that are necessary for learning semantic and procedural information and creating pointing gestures (2.5).

## 2.1 Ubiquitous Computing and Smart Domestic Environments

The ultimate purpose of room-based interaction is to offer selection techniques for smart environments: these environments delineate the design space and working domain of my dissertation and inform many design decisions I made throughout my research. Although smart environments have been given many names in the past, such as Ubiquitous Computing (UbiComp), Pervasive Computing, Ambient Intelligence (AmI), Smart Environments, and the Internet of Things, their overarching idea is to digitally enhance every-day artifacts in order to support people's lives. In detail, however, these five types of smart environments take slightly different approaches.

Ubiquitous Computing, a term coined by Mark Weiser (Weiser, 1991), puts an emphasis on augmenting people's environments with "hundreds of computers [that] will come to be invisible to common awareness. People will simply use them unconsciously to accomplish everyday tasks." (*Ibid.*, p. 98). One of Weiser's key point was that computers should disappear from people's conscious focus of attention and just "be a tool through which [they] work" (*Ibid.*, p. 76). The everyday task should be in the center of people's attention, not the tool they use to complete the task. Despite this broad statement, all of Weiser's examples took place in office environments

Pervasive Computing is a superset of UbiComp and also includes the notions of mobility and transition between different (mobile and static) environments (Satyanarayanan, 2001). Both the terms Ambient Intelligence (AmI) and Smart Environment refer to the idea that the environment can create awareness about itself and the people within through the use of sensors. The term Smart Environments captures the broader notion of an environment "that is able to acquire and apply knowledge about [itself] and also to adapt to its inhabitants" (Cook and Das, 2004, p. 3), whereas AmI focuses more on the people within the environment, their goals, and their activities. Situation recognition and implicit interaction are two key elements that set AmI apart from other UbiCom-related approaches (Ducatel, Bogdanowicz, Scapolo, Leijten, and Burgelman, 2001). The term Internet of Things originated in supply chain management with the purpose of electronically tagging every real-world object to make it recognizable by the system (Brock, 2001). Later, this term was reshaped to express the idea that everyday objects are connected to a network and can thus be digitally accessed and programmed by users (Gershenfeld, Krikorian, and Cohen, 2004). Since this dissertation is about investigating interaction techniques, I will not get into more detail on UbiComp and smart environments, such as communication protocols, hardware and implementation details, and real-world deployment of UbiComp systems and technology.

### 2.1.1 Human-Computer Interaction in Smart Domestic Environments

In UbiComp, research tended to focus more on technological than human factors, and human-computer interaction with UbiComp systems was mostly technology-driven (Abowd, 2012). Similarly, in my literature review I mainly focused on technological and cognitive human factors of human-computer interaction. Some researchers argued, however, that social factors have to be more carefully considered in designing successful interactions for smart domestic environments than they are currently (Edwards and Grinter, 2001). Ethnographic research has shown that people's activities in domestic environments are in many regards different from their activities in the—predominantly studied—office environments. First, activities in offices are generally more goal-oriented ("get job done"), whereas activities in domestic environments are more maintenance-oriented ("keep kitchen clean") (O'Brien and Rodden, 1997). Second, activities in offices are generally more team-oriented, whereas activities in domestic environments are more individual-oriented due to the "highly disparate priorities of different family members" (Tolmie, Pycock, Diggins, MacLean, and Karsenty, 2002, p. 399). Third, activities in offices are generally

more “understood in terms of ‘tasks’” (Crabtree and Rodden, 2004, p. 194), whereas activities in domestic environment are thought of in terms “daily routines”, which are “mundane yet essential activities […] of householders ordering their lives” (O’Brien and Rodden, 1997). Routines play a crucial role in people’s daily life as “routines are the very glue of everyday life [and] provide the grounds whereby the business of home life gets done” (Tolmie et al., 2002, pp. 399–400). In other words, routines dominate our domestic life and, on a day-to-day base, are important for making us function and survive within society. These routines are oftentimes learnt from parents and have been rehearsed and internalized through years of practice and reinforcement. Technology tampering with people’s routines can therefore have tremendous ramifications on people’s lives.

### 2.1.2 Primary Tasks and Supporting Tasks

In the context of my dissertation, a primary (or main) task is an activity that people have to complete in order to reach a goal. Examples include daily routines, such as ironing cloths (in order to have neat-looking shirts), or leisurely activities, such as watching a TV show (in order to relax). A supporting task is an activity that does not serve the main goal but is required for completing the main task, for example, setting up the ironing board and turning on the lights in order to iron and turning on the TV and selecting the right show on Netflix. One of the main goals of smart environments is supporting people in completing these supporting tasks more efficiently, or how Weiser put it: “unconsciously […] accomplish[ing] everyday tasks” (Weiser, 1991, p. 98). For this, “Ubicomp devices […] must not interrupt or distract the user from performing a primary task” (Landay and Borriello, 2003, p. 94). This means that any technique for Human-Environment Interaction should minimize the cognitive and temporal demand on the user.

## 2.2 Pointing-based Selection Mechanisms, Awareness, and Selection Proxies

In this section, I first define some selection technique related terms that I will use throughout my dissertation (2.2.1). I then present the history of HCI research on mid-air full-arm pointing gestures as selection mechanism (2.2.2) and give details about the current state (2.2.3). After this I describe how gestural interfaces can be used to create awareness (2.2.4). Finally, I explain the dual role of selection proxies as interaction facilitators and artifact representations (2.2.5).

### 2.2.1 Anatomy of a Selection Technique

Interactions with a computer system involves multiple conceptually different components. The domain object is the data or digital artifact that people set out to manipulate by modifying its attributes. The instruments are the mediator that converts users' actions into system commands, which then alter the domain object. There are two types of instruments: the mechanisms, which are the physical input devices and the actions they enable, and the proxies, which are the digital representations of the interaction within the user interface (Beaudouin-Lafon, 2000). Since my dissertation focuses on selections in smart environments, I will use the terms **selection mechanism** and **selection proxy** throughout. In room-based interaction, for example, pointing gestures act as *selection mechanisms* and real-world proxy objects act as *selection proxies*.

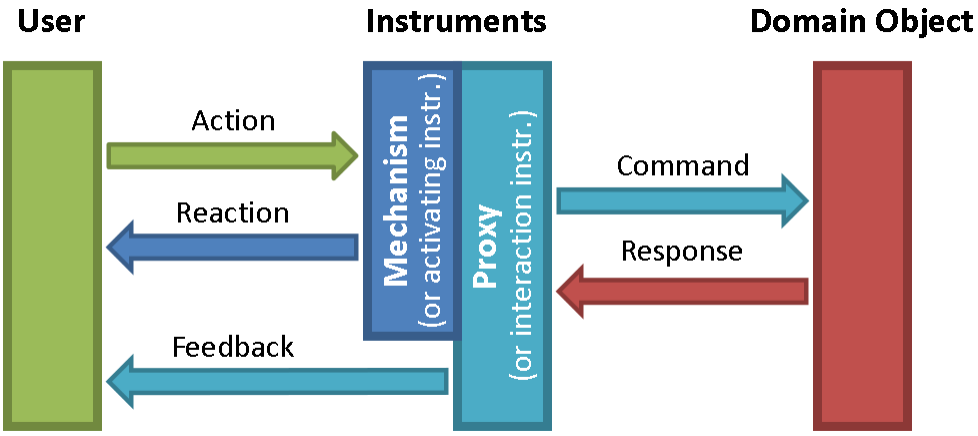



**Figure 1: Interaction model for post-WIMP interfaces; adapted from (Beaudouin-Lafon, 2000)**

### 2.2.2 A Brief History of Pointing Devices and Graphical User Interfaces

When talking about pointing devices or using pointing as selection mechanism in computer systems, indirect proxy-based devices—for example, mice—come into one's mind. Yet, the earliest device of this particular category was the trackball, first built in 1953 to control the graphical user interface (GUI) of a computerized battlefield information system called Digital Automated Tracking and Resolving (Vardalas, 1994). The first device in the shape of today's mice was designed by Douglas Engelbart in 1963 and filed for patent in 1967 (Engelbart, 1970); the first mouse that use a rolling ball, which would be the standard technology for over thirty years, was contrived in 1968 by a now defunct German company called Telefunken (Computer

History Museum, 2015a); the first optical mouse, which we are still using today, was created by Richard Lyon at Xerox (Lyon, 1981). The Xerox Alto from 1973 was the first example of what is now considered to be a personal computer (Computer History Museum, 2015b). It used a mouse to control the user interface (UI); this UI also was the first graphical user interface (GUI) that followed the windows, icons, menus, and pointer- (WIMP) metaphor. Yet, further innovations in input devices and interaction techniques arrived only slowly until other forms of computer systems besides terminals and desktop computers emerged.

### *From Graphical User Interfaces to Alternate Modes of HCI – 1980*

In 1980, Bolt presented a computer system with a UI that used a combination of speech and gestures to create objects on a wall-sized display (Bolt, 1980). Five years later, Krueger et al. built *VIDEOPLACE*, a prototype consisting of digital walls and desks that used arm movement to interact with wall-sized displays and digital tabletops (Krueger, Gionfriddo, and Hinrichsen, 1985). The authors intended their system to be used as an art installation for exploring "alternate modes of human-machine interaction" (*Ibid.*, p. 35); nonetheless, the system already exhibited many features of today's interactive tables, such as mid-air gesture recognition, embodiments, and touch-interaction. With *Charade*, Baudel and Beaudouin-Lafon presented a more elaborate version of Bolt's original idea in (Baudel and Beaudouin-Lafon, 1993). Through pointing with their hands, users could select so-called "active zones" (*Ibid.*, p. 30) on a wall-sized display and then interact with them through hand gestures. The authors also established two different paradigms that evolved from Bolt's original work: the *manipulation* and the *sign-language paradigms*. Through the manipulation paradigm, users interact directly with (pseudo-) physical objects; the sign-language paradigm lets users "issue commands with hand gestures" (*Ibid.*, p. 28). The manipulation paradigm has its roots in mixed reality systems (Appino, Lewis, Koved, Ling, Rabenhorst, and Codella, 1992) but was later applied to other user interface types, such as tangible user interfaces (TUIs) (Ishii and Ullmer, 1997), ubiquitous computing (UbiComp) (Fitzmaurice, 1993), and augmented realities (ARs) (Rekimoto and Nagao, 1995). From now on, I will refer to gestural techniques following the manipulation paradigm as **manipulation-based gestures** and to techniques following the sign-language paradigm as **sign-based gestures**.

### *Mixed Realities (MRs) – 1968*

Mixed realities (MRs) fill the "Virtuality Continuum" between real and virtual environments, former "consisting solely of real objects" and latter "consisting solely of virtual objects" (Milgram and Kishino, 1994, p. 1321); augmented realities (ARs), for example, lie within this spectrum, slightly leaning toward real environments. Azuma defines ARs as systems that "[allow] the user to see the real world, with virtual objects superimposed upon or composited with the real world", whereas virtual environments "completely immerse a user inside a synthetic environment" (Azuma, 1997, p. 356). Virtual environments (VEs), or virtual realities as they are more commonly called, completely immerse users inside a synthetic environment. While immersed, users cannot see the real world around them. In contrast, ARs allow users to see the real world, with virtual objects superimposed upon or composited with the real world. Therefore, ARs supplement reality, rather than completely replacing it. The history of MRs reaches back into the 1960 when MR-pioneer and Turing Award laureate Ivan Sutherland presented the first head-mounted display (Sutherland, 1968).

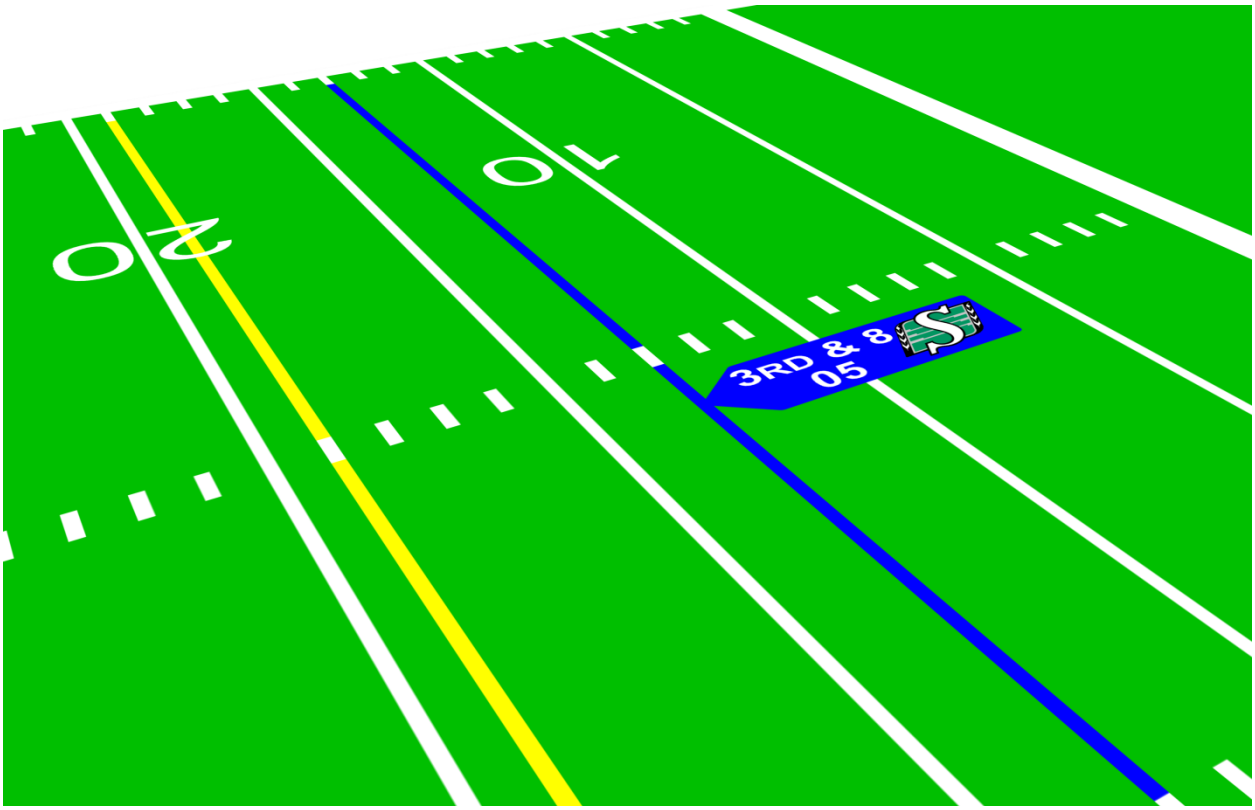



**Figure 2: Example of augmented reality in television: line to gain (yellow line), the line of scrimmage (blue line), number of the current down (3rd), the distance to a new 1st down (8), current possession (Saskatchewan Roughriders), and the play clock (05)**

Early research focused mainly on developing and improving new hardware displays and interaction device technologies. While display technology is of minor interest here, I will give a short overview of interaction devices. Since MRs require users to make input based on a three-dimensional environment, the first MR input devices were derived from 3D-input for GUIs, such as the *Lincoln Wand* (Roberts, 1966). The *Lincoln Wand* is the first example of using ultrasonic sound to track position of objects in three dimensions; the *Sketchpad* by Sutherland, in contrast, used visible light (Sutherland, 1968). Burton and Sutherland were the first to mention the necessity of using three-dimensional input for MRs: "It seemed obvious at first that corresponding three-dimensional computer input devices might be interesting and useful" (Burton and Sutherland, 1974, p. 513). Given the hardware-centric development of MRs, it is difficult to pinpoint when scientists started to address human-factor-related issues of these systems with their research. One early paper that touched upon these issues was the aforementioned *VIDEOPLACE* (Krueger, Gionfriddo, and Hinrichsen, 1985). The authors also foreshadowed the fields of Ubiquitous Computing and Tangible User Interfaces by demanding that "the human-machine interface is [to be] generalized beyond traditional control devices to permit physical participation with graphic images" (*Ibid.*, p. 35).

### *Ubiquitous Computing (UbiComp) – 1991*

In order to push the development of personal computers in a different direction, Weiser suggested in 1991 to integrate dedicated and visible computers into the environment, thus making them invisible. For Weiser, the strong physical presence of previous UIs made them failures in "making computing an integral, invisible part of people's lives" (Weiser, 1991, p. 94). To overcome this shortcoming, he suggested integrating computers into the world and thus "invisibly enhancing the world that already exists" (*Ibid.*, p. 94). He believed that "only when things disappear in this way are we freed to use them without thinking and so to focus beyond them on new goals" (*Ibid.*, p. 94). Weiser assumed that the disappearance of computers would ultimately have "a fundamental consequence not of technology but of human psychology" (*Ibid.*, p. 94). Finally, he argued strongly against VR on the basis that it "it excludes […] the infinite richness of the universe" (*Ibid.*, p. 94). For him, these systems, "which attempt to make a world inside the computer", are "most diametrically opposed to our vision [UbiComp]" (*Ibid.*, p. 94).

Fundamentally, UbiComp does not demand a certain way of interacting with all the integrated computers in intelligent environments, although it can be argued that their invisibility favors some and excludes other means of interaction. Ultimately, Weiser believed that the ubiquity of digital devices will have a profound social and economic influence on our lives—an effect that was undoubtedly achieved with the rise of smart phones.

### *Tangible User Interfaces (TUIs) – 1995*

One interaction metaphor that is directly rooted in the idea of UbiComp is the one of tangible user interfaces. The idea behind TUIs is to "allow users to 'grasp & manipulate' bits in the center of users' attention by coupling the bits with everyday physical objects and architectural surfaces" (Ishii and Ullmer, 1997, p. 234). Most TUIs therefore share a certain set of common characteristics, such as interaction through direct touch and use of real-world objects as interaction proxies. The idea of proxies makes TUIs of particular interest in the context of this research. Fitzmaurice et al. created *Bricks*, the first interaction technique that one could consider to be a TUI (Fitzmaurice, Ishii, and Buxton, 1995). Although called a graspable interface at that time, Bricks already contained the idea of TUIs because "bricks are essentially new input devices that can be tightly coupled or 'attached' to virtual objects for manipulation or for expressing action" (*Ibid.*, p. 442).


**Figure 3: *Tangible Bits* (left) (Ishii and Ullmer, 1997) and *Reactable* by Reactable Systems, SL (right)**

### 2.2.3 Manipulation-based Full-arm Pointing Techniques

The core idea of manipulation-based pointing techniques is that interaction with digital systems occurs through manipulation (primarily selection) of a physical or pseudo-physical proxy object (see 2.2.5 for more details about proxies). A classic example is the WIMP paradigm (Clarke, 1986), in which people use a pointer (e.g., a cursor) to manipulate on-screen proxy objects, such as icons and menu entries. Counterexamples are non-WIMP interfaces, such as traditional command-line interfaces, or post-WIMP interfaces, which have been discussed since the early 1990 (Beaudouin-Lafon, Ravn, Ratzer, et al. 2001; Beaudouin-Lafon, 2004). Sign-based gestures are oftentimes used in post-WIMP interfaces, for example, in *Wear Ur World* (Mistry, Maes, and Chang, 2009) and *Imaginary Interfaces* (Gustafson, Bierwirth, and Baudisch, 2010); see (LaViola, 2014) for an overview. The term 'pointing' is used rather broadly in HCI; for example, using a cursor is considered pointing although users are not performing an actual physical pointing gesture. In this work, I focus on distal pointing, which means that people use their whole arm to point at a distant real-world objects within a physical environment. For more detail on non-verbal communication and a full definition of distal pointing, please refer to sections 2.3.1 and 2.3.2.

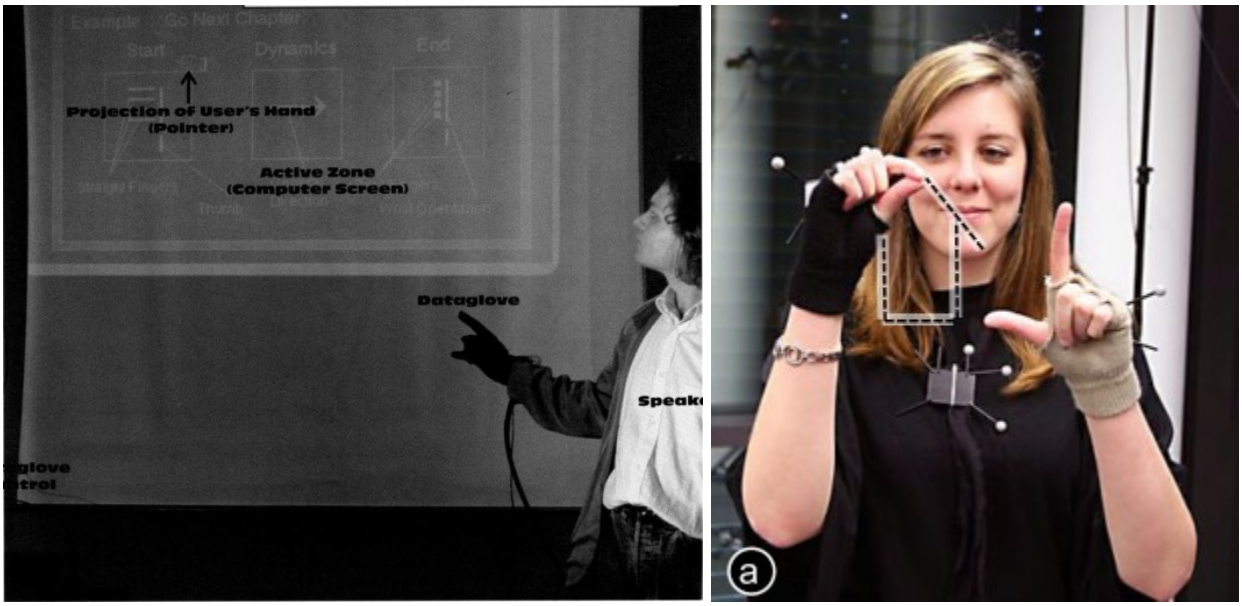



**Figure 4: *Charade*, a manipulation-based full-arm pointing technique (left) (Baudel and Beaudouin-Lafon, 1993) and *Imaginary Interfaces*, a sign-based full-arm pointing technique (right) (Gustafson, Bierwirth, and Baudisch, 2010)**

### *Distal Pointing in HCI*

Several HCI systems have been developed that use distal pointing for interacting with digital artifacts at a distance. *Charade* was the first system that used manipulation-based full-arm pointing gestures to interact with computers (Baudel and Beaudouin-Lafon, 1993). Users had to wear a data glove, which was tracked by the system, and were able to "issue commands by pointing at the active zone and performing gestures" (*Ibid.*, p. 30). These active zones were displayed on a wall-sized screen. *Charade* therefore used both manipulation-based gestures (for proxy selection) and sign-based gestures (for selection confirmation). The authors focused mainly on the evaluation of their gesture library and not on pointing performance. Independently, Marrin built the so-called *Digital Baton*, a location- and orientation-aware piece of hardware that allowed users to interact with a digital system through movement patterns and hand pressure (Marrin, 1997). Since this device is based on a baton, the author did not consider pointing as a mode of interaction.

Wilson and Shafer saw their *XWand* as an extension of the *Digital Baton* and explicitly included full-arm pointing gestures as selection mechanism (Wilson and Shafer, 2003). The *XWand* used a magnetic sensor and a gyroscope to determine its orientation and active infra-red-based camera-tracking to determine its location in space. Users could make selections by pointing the *XWand* at IR-reflective markers in the environment. The authors implemented target detection as a thresholded Gaussian probability distribution that calculated the likelihood with which the pointing ray intersects a target. Internally, targets were represented by a point cloud. While their paper focused on hardware details, they did run a short user study to assess users' pointing performance. The authors found that users' location within the environment did not affect pointing accuracy or completion time, although it has to be mentioned that users were facing the targets in all four possible starting locations (depending on condition: accuracy between 80% and 90%, completion time between 5.2 $s$ and 6.9 $s$). Wilson and Shafer envisioned the *XWand* primarily to be a selection tool for hardware devices, such as lamps. Users would execute a selection by pointing at a target and holding the *XWand* still for a short time. Users would then issue the actual command to the device through a different interaction modality, such as speech or (sign-based) gestural input.

In the same year, Wilson and Pham introduced the *World Cursor* to complement the *XWand* (Wilson and Pham, 2003). The *World Cursor* addressed the problem of imprecise system feedback about where users are currently pointing by adding a ceiling-mounted laser-pointer that would indicate the current pointing target. Users could either select real-world objects by pointing at them, which the authors called "absolute pointing" (*Ibid.*, p. 498), or by steering the laser pointer with the *XWand*, which was dubbed "relative pointing" (*Ibid.*, p. 498)). The reasons for the authors to include relative pointing were hardware-related, in particular calibration, line-of-sight, and precision issues. Despite the fundamental difference between these two pointing paradigms, the authors did not evaluate user performance or preference.

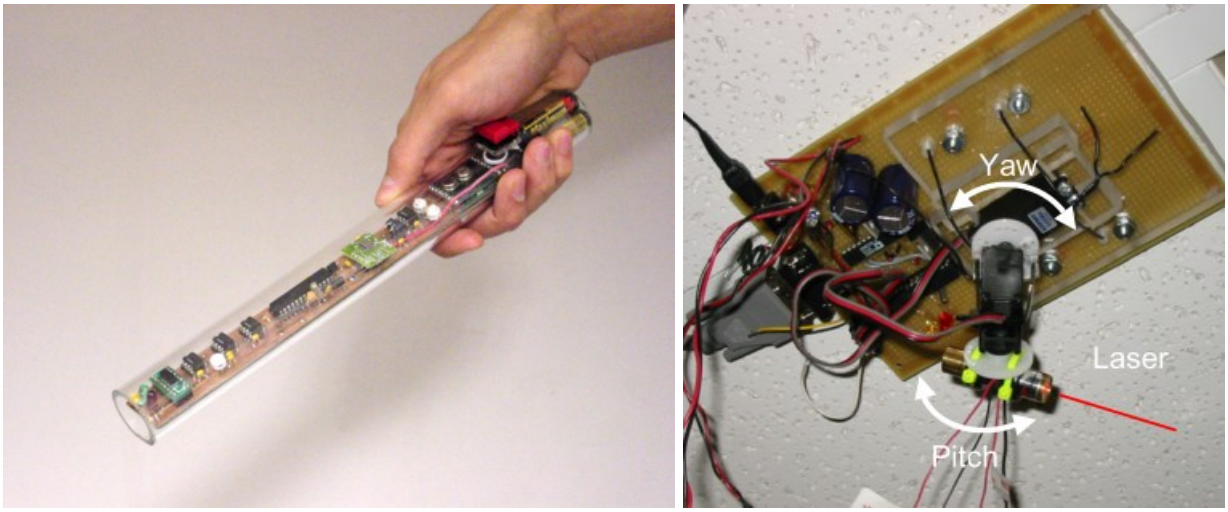



**Figure 5: *XWand* (left) (Wilson and Shafer, 2003) and *World Cursor* (right) (Wilson and Pham, 2003)**

With the *FindIt Flashlight*, Ma and Paradiso presented a system that uses absolute pointing at photo-sensitive tags for selection (Ma and Paradiso, 2002). The authors used pulse-coded optical transmission in order to avoid interference from natural light; because of this design choice, the system could potentially distinguish between different users. Patel and Abowd later used the same system, but replaced the broad flashlight beam with a—spatially more coherent—laser beam (Patel and Abowd, 2003). Since both project were hardware proof-of-concepts, none of them included a user evaluation.

The *VisionWand* project by Cao and Balakrishnan built upon the *XWand* and extended its gesture alphabet (Cao and Balakrishnan, 2003). The initial purpose for the *VisionWand* was selection and manipulation of objects on a wall-sized display. Distal pointing retains its role as selection mechanism, whereas other sign-based gestures, such as tilting and rotating, are used for object manipulation. As with the *XWand*, the authors were more interested in a hardware proof-of-concept and did not conduct a user study to assess the usability of their interaction technique.

Vogel and Balakrishnan were amongst the first who evaluated users' distal pointing performance in selection tasks (Vogel and Balakrishnan, 2006). They compared three selection techniques (relative clutching, ray-to-clutching, and ray-casting) for wall-sized displays. Overall, the authors found that ray-casting required the lowest selection time while having the highest error rate, especially for smaller targets. For large targets ($144\ mm$ diameter), all three techniques showed the same error rate.

### *Fitts's Law*

*Fitts's Law* (Fitts, 1954) models human movement as transmission of information (MacKenzie, 1992). It is now widely used in HCI for predicting human performance as it "has proven [to be] one of the most robust, highly cited, and widely adopted models to emerge from experimental psychology" (*Ibid.*, p. 93). However, the original version of Fitts's Law, $ID = \log_2(A/W + 1)$, cannot be directly applied to distal pointing tasks because it does not take the distance between user and target into consideration (Kopper, Bowman, Silva, and McMahan, 2010).

There have been multiple attempts to adapt Fitts's Law to distal pointing. Kopper et al. extended Fitts's Law such that $ID_{angular} = \left[\log_2(\alpha/\varpi^k + 1)\right]^2$, where $\alpha$ is the angular amplitude of the movement ($\alpha = 2\arctan A/2D$), $\varpi$ the angular width of the target ($\varpi = \arctan A + W/2D - \arctan A - W/2D$), and $k$ a power factor that determines the relative weights of $\alpha$ and $\varpi$ (Kopper, Bowman, Silva, and McMahan, 2010, pp. 606, 611). The authors conducted regression analyses on multiple data sets that they collected and showed that their model fits the data well ($R^2$ between $.945$ and $.961$).

Soechtig and Lacquaniti investigated the spatial and temporal characteristics of people's pointing gestures during a Fitts-style task (Soechting and Lacquaniti, 1981). They found that "motion at

the shoulder and elbow is tightly coupled" (*Ibid*, p. 130), while "motion at the wrist is much more variable in relation to motion at the more proximal joints" (*Ibid.*, p. 130). Furthermore, they showed that people use shoulder- and elbow-movements for an initial approach to a pointing target, whereas the wrist is responsible for acquiring the target once the hand gets close.

Lastly, Oh and Stuerzlinger applied Fitts's Law to a comparison of laser pointers with traditional mice using an ISO standard pointing task (Oh and Stuerzlinger, 2002). They concluded that "the average throughput of the laser pointer is about 75% of the mouse" (*Ibid.*, p. 6). Furthermore, they found that completion time was significantly higher when using laser pointers.

### *Laser Pointers as Input Devices*

Myers et al. experimented with different laser pointer designs, ranging from a standard cylindrical laser pointer to a gun-mounted laser pointer (Myers, Bhatnagar, Nichols, Peck, Kong, Miller, and Long, 2002). They found that designs have a significant influence on pointing accuracy; yet, the average angular error between designs at $4.5\ m$ distance did not exceed $0.18°$. Lastly, the authors confirmed Oh and Stuerzlinger's finding that mice have roughly 50% more throughput than laser pointers (Oh and Stuerzlinger, 2002). Kemp et al. evaluated the use of laser pointers in human-robot interaction (Kemp, Anderson, Nguyen, Trevor, and Xu, 2008). With their system, the authors wanted to enable wheelchair-bound people to use laser pointers to communicate objects of interest to an autonomous robotic system for retrieval. The system was able to detect laser pointers with an angular error of 4.8° or less, which can be partially attributed to both pointing imprecision and detection inaccuracies. Although it remains unclear whether pointing imprecision or detection inaccuracies were the source for the pointing errors, the results give another upper limit for human pointing performance with laser pointers.

Some issues with laser pointer interaction, such as jitter, detection error, slow sampling, and latency, were mentioned early on, for example, by Olson and Nielsen (Olsen and Nielsen, 2001). Since muscular jitter is an unavoidable human issue, it is still relevant for today's interaction designer. Another important issue is the involuntary pitch- or yaw-movement of a tracked device that can occur when a user is pressing the on-device button. Bowman et al. dubbed this the *Heisenberg Effect*, a "phenomenon that on a tracked device, a discrete input (e.g. button press) will often disturb the position of the tracker" (Bowman, Wingrave, Campbell, Ly, and Rhoton, 2002, p. 124). Segen and Kumar quantified the amount of location-based and orientation-based

jitter for their sign-language-based interaction technique. They concluded that jitter does not exceed 5 $mm$ (location) and 2° (orientation) (Segen and Kumar, 1998). The human visual system also sets a lower boundary for the size of targets, although this limit is lower than the one set by the human motor system (see 2.4.2 for more details on the human sensory system). Hourcade and Bullock-Rest conducted a study on a traditional WIMP-based interface and showed a strong increase in users' selection time and error rate when the angular size of the visible target fell below 0.05° (Hourcade and Bullock-Rest, 2012).

### *Hybrid Techniques: the Continuum between Manipulation-based and Sign-based*

Some interaction techniques use pointing gestures in a way that users can interpret these gestures both in a manipulation-based and in a sign-based way depending on the users' mental model.

Li et al. introduced *Virtual Shelves*, a selection technique that partitions the body-relative space around users in invisible segments, to which users can assign digital items (Li, Dearman, and Truong, 2009). Distal pointing at one of these segments invokes the associated command. The authors conducted a user study, in which they measured participants' pointing accuracy and selection time. They reported that users could fairly accurately point horizontally but not vertically; in general, user performance was worse than reported in previous research. Selection time was similar in both directions (3.09 $s$ horizontally, 3.14 $s$ vertically). Furthermore, the authors conducted a short study in which they mapped 28 digital items to *Virtual Shelves*. A rough evaluation showed that *Virtual Shelves* is indeed faster than a traditional phone interface, which was used as comparison, but also produced more than four times as many selection errors.

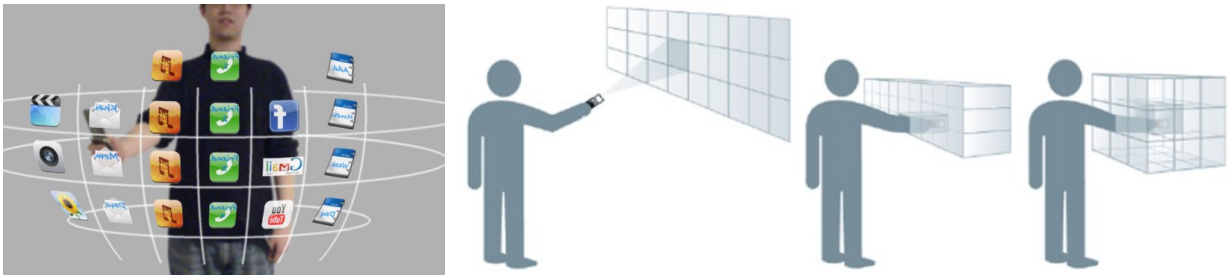


**Figure 6: *Virtual Shelves* (left) (Li, Dearman, and Truong, 2009) and *Air Pointing* (right) (Cockburn et al., 2011)**

Cockburn et al. conducted a more thorough study of a class of selection techniques called *Air Pointing* (Cockburn, Quinn, Gutwin, Ramos, and Looser, 2011). *Ray-casting* is the selection technique comparable to *Virtual Shelves*. A major difference was that participants were only allowed to move their wrists instead of their whole arm and that there were only four selection targets. The authors reported a substantially lower completion time than Li et al. of just below $1\ s$; pointing accuracy was high when feedback was given (around 3°) but worsened substantially without feedback (to around 18°). The authors attributed this drop in accuracy to drifting, an effect in which a participant's pointing gesture gradually drifted away from the actual target. Besides the user evaluation, the authors created a framework for the classification of pointing-based selection techniques, which could be a useful tool for a systematic design space investigation (see below).

Researchers oftentimes use optical motion tracking systems, such as the Vicon *Bonita* (Vicon Bonita, 2015) or the NaturalPoint *OptiTrack Prime* (NaturalPoint OptiTrack Prime, 2015), which are too expensive for an average household. Thus, one might argue that we should not research human factors for a selection method that requires overly expensive hardware. There are two examples of more recent projects that capture people's full-arm pointing gestures with systems more likely already installed in people's homes. First, Dutta showed that the Microsoft Kinect is useful for capturing objects up to a precision of $1\ cm$ (Dutta, 2012); accuracy can be further improved using input from multiple Kinect sensors (Caon, Yue, Tscherrig, Mugellini, and Khaled, 2011). Second, researchers have begun exploiting existing the electromagnetic field generated by power lines, household electronics, and wireless networks (Cohn, Morris, Patel, and Tan, 2012; Pu, Gupta, Gollakota, and Patel, 2013). Any human motion within this field causes changes of it; these changes can be detected by simple antennas. While neither of these two novel approaches have yet been tested in the context of accuracy of full-arm pointing gestures, they show that affordable motion capturing might indeed be feasible in the near future.

### *Taxonomies of Manipulation-based Pointing Interaction*

It is difficult to define a unified taxonomy of all pointing input devices or pointing interaction techniques. Instead, researchers have come up with many different taxonomies that approach the field of HCI from different perspectives and by looking at different cross-cutting issues, for

example, output modalities (Bernsen, 1997), sub-areas, for example, virtual reality systems (Coomans and Timmermans, 1997), and tasks, for example, research (Agah, 2000).

One taxonomy that is useful for this research was drafted by Karam and Schaefel to characterize gestural interaction (Karam and Schraefel, 2005). Although their work was not published in a peer-reviewed journal, it is informative because the authors combined their taxonomy with a very thorough literature review from psychology, linguistics, and computer science. The taxonomy consists of four dimensions: application domain, enabling technology, system response, and gesture style. The "gesture style" dimension is particularly interesting since it follows a psychology-based approach to gestures (*Ibid.*, p. 3). The authors based this dimension on earlier work in linguistics by Efron (Efron, 1941) and McNeill (McNeill, 1992). I go into more detail about gestures and non-verbal behavior in section 2.3.1.

A second framework useful for classifying full-arm pointing gestures is the *Air-pointing Design Framework* by Cockburn et al. (Cockburn, Quinn, Gutwin, Ramos, and Looser, 2011). The authors refer to selection gestures that require "moving a limb, finger, or device to a specific spatial region" (*Ibid.*, p. 401) as *Air Pointing*. I argue for extending the framework's narrow scope beyond the domain of the gestures described above, to include full-arm pointing gestures as well, for two reasons. First, moving and pointing are just two different interpretations of similar limb movements; second, Cockburn et al. themselves used a pointing technique (*Ray-casting Air-pointing*) within their framework. Since the *Air-pointing Design Framework* allows precise definitions of new interaction techniques and positioning them in relation to existing techniques, it is very useful for setting the context of this research.

### *The Air-pointing Design Framework*

The Air-pointing Design Framework (Cockburn, Quinn, Gutwin, Ramos, and Looser, 2011) defines five interaction dimensions:

- reference frame for spatial input,
- scale of spatial input control,
- degrees of freedom in spatial input,
- feedback modality, and
- feedback content.

*Reference frame* defines the origin of the coordinate system in which the spatial input takes place. The origin can either be fixed to the environment (absolute location), a movable object (object-relative), the interaction device (device-relative), or to the user (body-relative). *Input scale* defines the size of the "physical movements for controlling spatial input" (*Ibid.*, p. 406) and can be as small as a finger twitch or as large as a full-body movement. *Input degrees of freedom* define which of the six spatial dimensions are interpreted as user input. *Feedback modality* describes the sensory channels through which users receive feedback about their spatial input. Finally, *feedback content* describes in more detail what feedback information is sent to users.

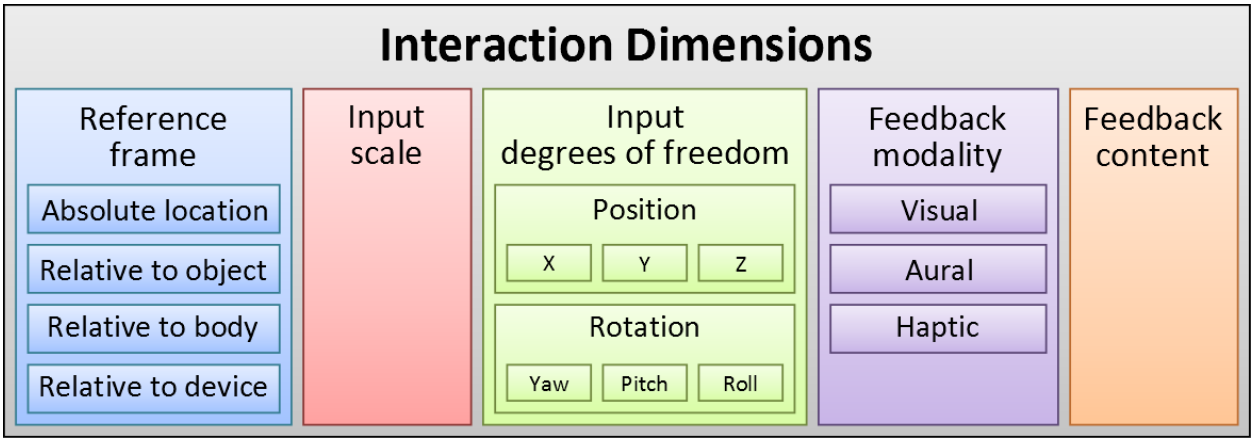



**Figure 7: *Air-pointing Design Framework* (Cockburn et al., 2011, p. 405)**

A sign-based gesture like in *Imaginary Interfaces* (Gustafson, Bierwirth, and Baudisch, 2010), for example, can be characterized using this framework as body-relative, small input scale, 6 degree of freedom, with visual and proprioceptive feedback. In contrast, a manipulation-based gesture like in *XWand* (Wilson and Shafer, 2003) can be characterized as absolute, large input scale, 6 degree-of-freedom, with visual and proprioceptive feedback.

For the purpose of this research, I am going to focus on interaction techniques with manipulation-based full-arm pointing gestures (see 2.2.3) and interaction techniques that use static real-world objects as proxies (see 2.2.5). These two constraints nicely define the design space I am interested in. By using static real-world entities as proxies, the reference frame is set to absolute location; by using full-arm pointing gestures, input scale is set to full arm; by using

pointing in combination with real-world entities, the input degrees of freedom are set to $X, Y, Z, Yaw, Pitch$; and by using full-arm pointing, the feedback modalities are set to visual and proprioceptive.

### 2.2.4 Awareness of People and Their Actions

Awareness is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988, p. 97). In other words, it is the knowledge about the current state of things in the environment, how their state will change, and how these changes impact oneself. The information required for creating awareness is transmitted through a communication medium, for example, sounds, vision, or text (Benford and Fahlén, 1993). The awareness of other people and their action is an important aspect of people's daily life. Situation awareness in traffic, for example, is essential to prevent accidents and harm (Endsley, 1988), and group awareness in a shared office space is essential to collaborate efficiently with co-workers (Gutwin and Greenberg, 2002).

#### *Groups, Coupling, and Mixed-focus Collaboration*

Groups are sets of two to five people who carry out tasks in medium-sized workspaces (Gutwin and Greenberg, 2002). Whenever people engage in collaborative activities, they have to split their attention between two tasks: their actual work (a primary task) and awareness maintenance (a supporting task) (Gutwin and Greenberg, 1998). People have to perform these two tasks concurrently, which means that these two tasks compete for people's attention. Subsequently, groupware designers try to minimize the cognitive load from awareness maintenance so that people are as little as possible distracted from their actual work.

Coupling refers to the degree with which people have to interact to progress with their work (Segal, 1994). The level can reach from individual work over loosely-coupled to tightly-coupled (Streitz, Haake, Hannemann, Lemke, Schuler, Schütt, and Thüring, 1992). The level of coupling also determines "the level of awareness each [person] has of the activity [of other collaborators]" (*Ibid.*, p. 13). For individual work, every person works on an individual task; for loosely-coupled work, multiple collaborators are "working on the same subtask [and] manipulate the same [artifact]"; for tightly-coupled work, collaborators have to "cooperate and coordinate their work in synchronous conference-like 'meetings'" (*Ibid.*, p. 13).

In group activities, people often engage in mixed-focus collaboration, i.e. people shift frequently between loosely and tightly coupled activities during a work session (Dourish and Bellotti, 1992). When people are loosely coupled, they have to interact less with each other to complete their task as when they are tightly coupled (Segal, 1994). However, even during loosely coupled work, people still need to be aware of others' activities (Rico and Brewster, 2010).

### *Group Awareness and Consequential Communication*

Group awareness is the understanding of the activities of others. It provides context for people's activities and is critical to successful collaboration (Dourish and Bellotti, 1992). Two factors determine the level of group awareness: the actor's nimbus and the observer's focus. Nimbus is the space within a communication medium in which acting people make their activity observable to others (dashed lines in Figure 8); focus is the space within a communication medium that is covered by the observers' attention (dotted lines in Figure 8) (Benford and Fahlén, 1993). "The level of awareness that object $A$ has of object $B$ in medium $M$ is some function of $A's$ focus on $B$ in $M$ and $B's$ nimbus on $A$ in $M$" (*Ibid.*, p. 112). There are three possible levels of awareness between actor and observers. When focus and nimbus of two people overlap, there is full awareness between them; when the focus of one person overlaps the nimbus of another, there is semi-awareness between them; and when the focus of the two people do not overlap each other's nimbus, there is no awareness. Figure 8 shows these three possibilities. When nimbus and focus overlap in stages of semi- or full awareness, observers go through a three-phase process to gain group awareness: perception of an action, comprehension of the situation, and projection of the future status (Endsley, 1995).

Control over artifacts determines the available mediums that can be used for creating awareness. When an artifact is under control of a single actor, observers have to rely on direct communication from the actor to create awareness about the state of the artifact (see Figure 9, left; Dix, 1994). Direct communication is explicit, oftentimes occurs through speech or gesture, and its sole purpose is awareness creation (Gutwin and Greenberg, 1996). When an artifact is shared, in contrast, "that artifact is not only the *subject* of communication, it can also become a *medium* of communication" (Dix, 1994, p. 13).

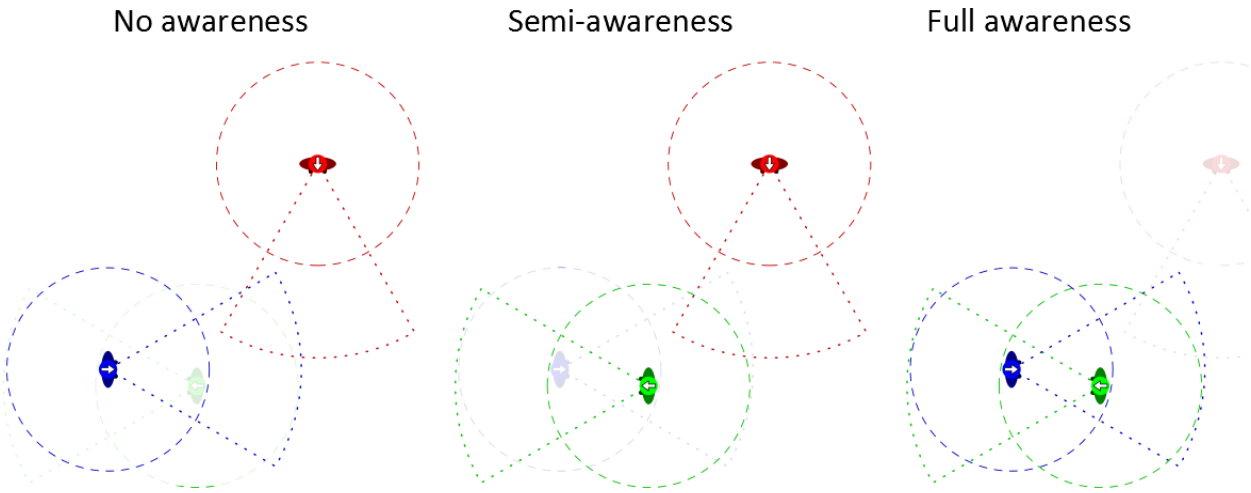



**Figure 8: Levels of awareness; dotted lines indicate a person's focus, dashed line the nimbus, white arrow the viewing direction (adapted from Benford and Fahlén, 1993)**

This new capability of an artifact is called feedthrough, a reference to the term "feedback": in addition to feeding information about the artifact's status back to the actor, it feeds information through to all observers as well (see Figure 9, center). In addition, the shared nature of an artifact oftentimes allows observers not only to perceive changes in the artifact's status but also the action that manipulated the artifact. This implicit information gained from the observation of an action is called consequential communication. Consequential communication occurs through visible or audible signs of interaction with a workspace (Storey, Čubranić, and German, 2005). The size of the actions (or selection mechanism, see 2.2.1) necessary to operate controls makes actions public and creates situation awareness, which is important in many collaborative real-world tasks (*Ibid*., 2005). In HCI research, consequential communication (see Figure 9, right) is frequently mentioned as an awareness mechanism, and observational studies show that it is frequently used in real-world situations (Storey, Čubranić, and German, 2005). However, it is rarely explored in controlled studies and occasionally considered to be of little importance (Streeck, 1993). This is in contrast to other fields, which showed that consequential communication plays a crucial role throughout life, for example, as facilitator for learning through observation and imitation (Hanna and Meltzoff, 1993; Salvador, Scholtz, and Larson, 1996).

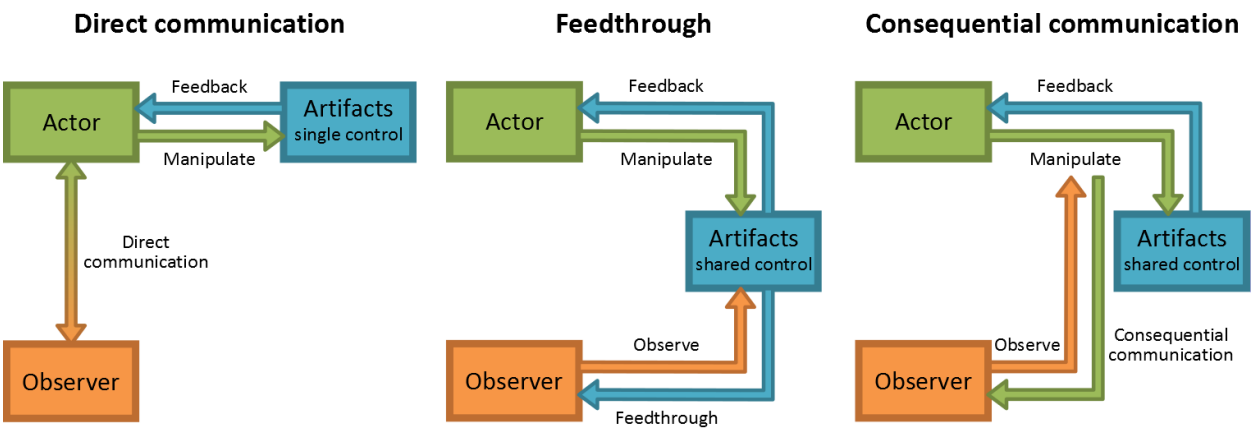



**Figure 9: Direct communication (left), feedthrough (center), and consequential communication (right) (adapted from Dix, 1994, p. 11-13, and Gutwin and Greenberg, 1998, p. 210)**

Besides direct communication, there are two more methods for creating group awareness in collocated environments: indirect productions and environmental feedback (Gutwin and Greenberg, 1996). Indirect productions are intentionally public but non-directed actions, which sets them apart from public and directed direct communication. Environmental feedback comes from the effect that a manipulation on a single artifact has on the environment (Gutwin and Greenberg, 1996). In this sense, environmental feedback is an indirect version of feedthrough.

In the context of my dissertation, consequential communication is of particular interest because the amount of awareness that observers can generate from consequential communication depends directly on how much information observers can extract from a selection mechanism.

### *Using Gestures for Awareness Creation*

The observation of other people's interaction with digital systems is closely related to the concepts of group awareness and consequential communication. The initial impulse in this area came from Don Norman in 1993 when he described the usefulness of "big controls and big actions" for shared work:

*The critical thing about doing shared tasks is to keep everyone informed about the complete state of things [...] each pilot or member of the control team must be fully aware of the situation, of what has happened, what is planned. And here is where those big controls come in handy. When*

*the captain reaches across the cockpit over to the first officer's side and lowers the landing-gear lever, the motion is obvious: the first officer can see it even without paying conscious attention. The motion not only controls the landing gear, but just as important, it acts as a natural communication between the two pilots, letting both know that the action has been done. […] Automatically, naturally, without any need for talking.*

— *Things that make us smart* (Don Norman, 1993, p. 142)

This kind of implicit information flow is called consequential communication and has been shown by several researchers to be an important part of the natural way in which people maintain awareness in a group (Carl Gutwin and Saul Greenberg, 1996; Leon Segal, 1994). However, consequential communication depends on large easily-observable actions and controls, which are no longer common in most workplaces. Instead, most tasks are now carried out on general-purpose computers with standard graphical user interfaces or on hand-held touch-devices. On these computers, activities that once had characteristic actions and artifacts (e.g., getting a file from a cabinet, using a Rolodex to find a telephone number, drawing a diagram, or entering numbers in a ledger) now all look very similar to an observer—that is, they all look like a person sitting at a computer monitor and moving a mouse or like a person holding a device and tapping on its screen.

Researchers in distributed groupware have looked at the problem of reduced observability (since people's bodies are not visible in a distributed setting), and have proposed visualization techniques to make others' actions in a shared workspace more obvious (Carl Gutwin and Saul Greenberg, 1998). However, these enhancements often work only when people are observing the same part of the shared workspace, and the techniques do not provide a solution in situations where people are carrying out loosely coupled work in a co-located setting.

### 2.2.5 Static Real-world Proxy-based Selection Techniques

Beaudouin-Lafon defines selection proxies as "mediators or two-way transducer between the user and the domain object" (Beaudouin-Lafon, 2000, p. 448). This definition stresses the dual role of proxies: they allow users to manipulate the digital artifact associated with the proxy and they convey information about the digital artifact to the user. In a WIMP interface, for example, an icon is a redirection to a data file. The icon provides a means for accessing the data file. With

a touch-screen, people can select the icon directly; on a traditional desktop computer, however, they need another proxy, such as a mouse cursor, to facilitate their selection request. The roles of proxies in these two examples are redirection and facilitation. The other role of proxies is that they also represent non-physical items. For example, when people want to access a data file in a WIMP interface, they can use the icon's spatial location and visual appearance to find it. In a similar fashion, the mouse cursor is a spatial and visual representation of the user's hand or finger. In the following sections, I discuss these two roles of proxy-objects.

### *Proxies as Redirection and Facilitation*

As mentioned above, interacting with digital systems is facilitated by selecting a proxy-object, such as an on-screen icon. For physical environments, which do not necessarily contain screens, interaction designers have to think about real-world proxies as alternatives for on-screen proxies. One interaction paradigm that makes heavily use of real-world proxies is TUIs. One of the first TUIs used *Passive Interface Props* by Hinckley et al. (Hinckley, Pausch, Goble, and Kassell, 1994). These props (puppet heads) were used to control the (perspective) viewport onto a three-dimensional skull model and thus played the role of an interaction proxy. The authors saw their props as a logical extension of on-screen icons into the third dimension. With *Bricks*, Fitzmaurice et al. turned the passive props into active manipulation tools for digital items. They saw their bricks as "new input devices that can be tightly coupled or 'attached' to virtual objects for manipulation or for expressing action" (Fitzmaurice, Ishii, and Buxton, 1995, p. 442). The authors' idea was to use real-world objects to select, move, rotate, and transform digital items in a drawing program on their *ActiveDesk*, which itself was an extension from Wellner's *DigitalDesk* (Wellner, 1993). The first attempt to associate real-world objects with digital items was *Tangible Bits* by Ishii and Ullmer (Ishii and Ullmer, 1997). Their so-called "phicons" are real-world objects, usually small and light enough to be picked up by hand, that serve as proxies for interacting with digital systems in the same way that icons do on WIMP-based GUIs. The authors' main argument for using real-world proxies is that "GUIs fall short of embracing the richness of human senses and skills people have developed through a lifetime of interaction with the physical world (*Ibid.*, p. 240).

*Tangible Bits* do not use pointing but direct touch as a selection method. Fitzmaurice's *Chameleon* (Fitzmaurice, 1993), on the other hand, uses pointing and can therefore be considered

a hybrid of Ishii's *Tangible Bits* (Ishii and Ullmer, 1997) and Baudel's *Charade* (Baudel and Beaudouin-Lafon, 1993). With *Chameleon*, Fitzmaurice introduced so-called "information hot spots on the physical device" (Fitzmaurice, 1993, p. 40). In contrast to a tangible interface, where users have to physically touch the proxy-object, they can now simply point at the "hot spot" and interact with the underlying information space (*Ibid.*, p. 40). Although the *Chameleon* system requires an UbiComp-enabled environment, the information-retrieval unit by itself could be considered an augmented reality device. This dual nature of TUIs is not surprising because, according to Wellner et al., the goal of augmented realities is to "create spaces in which everyday objects gain electronic properties without losing their familiar physical properties" (Wellner, Mackay, and Gold, 1993, p. 26).

After the initial idea of TUIs was published, researchers started exploring the interaction space of this new UI paradigm. Rekimioto and Saito included phicons into their *Augmented Surface* (Rekimoto and Saitoh, 1999). Since their system was set up around an interactive table, people could use phicons to display "the object aura [which] represents a data space for the corresponding object" (*Ibid.*, p. 381).

**Figure 10: *Passive Interface Props* (left) (Hinckley et al., 1994) and *Data Mountain* (right) (Robertson et al, 1998)**

With *Body Mnemonics*, Ängeslevä et al. took a slightly different approach and used body parts as selection proxies—for example, users could touch their shoulder to initiate a phone call (Ängeslevä, O'Modhrain, Oakley, and Hughes. 2003). Strachan et al. implemented a version of

*Body Mnemonics* called *BodySpace* , focusing primarily on hardware implementation issues (Strachan, Murray-Smith, and O'Modhrain. 2007). In contrast, Guerreiro et al. implemented a version of *Body Mnemonics* and conducted a short user evaluation (Guerreiro, Gamboa, and Jorge. 2007). In their study, participants were free to create their own association between a given set of commands and their body part. The authors reported that participants were able to achieve selection accuracies above 95 % with twelve mapped commands. The authors also found a strong correlation between certain commands (e.g., "SMS" and "Call") and different body locations (e.g., ear and hand)

### *Proxies as Representation*

The visual appearances of proxies and their spatil layout plays an important role when consdiering proxies as visual and spatial representation of data. Spatial layout in the context of finding items is closely related to the field of personal information management, which are "the activities a person performs in order to […] store [and] organize […] the information needed to complete tasks" (Jones, 2007, p. 453). Malone conducted a study in which he interviewed ten office workers about their organizing habits for (paper) files (Malone, 1983). From his observations he learnt that "the notion of accessing information on the basis of its spatial location, instead of its logical classification, is an important feature of the way people organize their desktops" (*Ibid.*, p. 108). He therefore suggested that location and color should be among the four properties that digital systems should support in order to aid users in finding files. However, not all studies agree that location is critical. One of the earliest evaluation of the use of spatiality in user interfaces was conducted by Dumais and Jones (Dumais and Jones, 1985). From their paper-based comparison of textual, spatial, and mixed interfaces they concluded that "location is neither an effective filing dimension in and of itself, nor does it appear to add much to the symbolic name" (*Ibid.*, p. 129). Still, the authors admitted that their study only acted as "a first attempt to assess the utility of a two-dimensional spatial representation" and that "performance might improve if the space were enriched through the introduction of landmarks" (*Ibid.*, p. 129). Dumais and Jones also hinted at the importance of context or meaning, an important tool for human memory (see section 2.5.5). Lansdale was one of the first authors who directly criticized the way in computer systems make data files available to users when he stated that "the process of information retrieval in the human mind is fundamentally different from filing […] systems" (Lansdale, 1988, p. 59). He stressed the importance of context, in which

people store information as a crucial part of the retrieval process since "our memory for detail is so much better if placed in the context of a wider scheme of things" (*Ibid.*, p. 59).

*Data Mountain* by Robertson et al. was one of the first attempts to store icons in a WIMP interface using three-dimensional spatial arrangements (Robertson, Czerwinski, Larson, Robbins, Thiel, and van Dantzich, 1998). In their study, the authors compared selection time and error rate between the Data Mountain UI (a simulated inclined plane on which icons could be placed) and a traditional bookmark menu. They concluded that "*Data Mountain* reliably facilitated speedy retrieval" and "users performed more accurately with […] *Data Mountain*" (*Ibid.*, p. 160). Cockburn and McKenzie implemented their own version of Data Mountain and compared it to a two-dimensional spatial arrangement, similar to a virtual desktop (Cockburn and McKenzie, 2001). They found no significant difference between these two conditions, and argued that the 2D- and 3D-visualization were very similar—and in fact, that the *Data Mountain* interface was not 3D but 2½D (i.e., 2D with overlap) (Cockburn and McKenzie, 2002). As a result, Cockburn and McKenzie conducted another study in which they compared 2D, 2½D, and 3D visualizations for icon selection on a virtual desktop as well as equivalent physical storage systems for object selection in the real world (*Ibid.*). For the virtual desktop, participants overall required least selection time for the 2½D visualization; in the real world, participants were fastest with a two-dimensional arrangement of objects.

Brumitt and Cadiz investigated what type of representation people preferred for controlling lights in a domestic setting. Given the choice between a floor plan on a touch-screen, a drop-down menu, a voice interface, a location-sensitive voice interface, and a combined voice and gesture interface, people preferred the voice and the combined voice and gesture interface and found them the easiest to use (Brumitt and Cadiz, 2001). Kühnel et al. conducted a study in which they asked participants to freely map 23 commands to gestures (Kühnel, Westermann, Hemmert, Kratz, Müller, and Möller, 2011). After this, the authors assigned the gestures into four categories according the gestures' nature: physical, metaphorical, abstract, and symbolic. In their analysis, the authors reported that physical gestures showed the highest agreement between participants as well as the lowest completion time.

Unfortunately, there is not much research about representational design aspects of real-world proxies in the area of TUIs. At this moment, researchers are still focusing on the technological

aspects of designing, tracking, and prototyping tangibles (Leitner and Haller, 2011) as well as augmenting everyday devices to make them communicate properly (Woo and Lim, 2012).

## 2.3 Pointing Gestures in Human–Human Communication

After presenting the current state of HCI research on using mid-air full-arm pointing gestures, I give a more rigorous definition of mid-air full-arm pointing gestures. I approach this definition from two directions: the purpose of different types of pointing gestures in human-human communication (2.3.1) and the distance between the acting person and the pointing target (2.3.2).

### 2.3.1 Types, Functions, and Purpose of Pointing Gestures

Pointing gestures are part of human non-verbal behavior. There are multiple ways for classifying non-verbal behavior; I use Ekman and Friesen's classification (Ekman and Friesen, 1981) because it subdivides gestures more finely than other classifications (McNeill, 1992). In 1981, Ekman and Friesen formalized an early classification by Efron (Efron, 1941) and distinguished between five types of non-verbal human behavior: emblems, illustrators, regulators, affect displays, and adaptors (Ekman and Friesen, 1981, p. 102).

*Emblems* are "nonverbal acts which have a direct verbal translation" (Ekman and Friesen, 1981, p. 71). They "occur most frequently where verbal exchange is prevented by noise, external circumstances, distance, by agreement, or by organic impairment. In such instances, emblematic exchange carries the bulk of messages which would typically be communicated through words" (*Ibid.*, p. 72). In summary, people use emblems as a substitute for words when verbal communication is impossible. *Illustrators*, in contrast, are "movements which are directly tied to speech, serving to illustrate what is being said verbally" (*Ibid.*, p. 76). *Deictic and spatial movements* (both sub-categories of illustrators) occur when the movement is "pointing to a present object" or "depicting a spatial relationship" (*Ibid.*, p 76). From these two definitions, the major difference between emblems and illustrators is the component of speech, which is absent in the first but required for the latter. In the context of HCI, pointing gestures can come from either category. Sign-based gestures (see 2.2.2 for the definition and *Imaginary Interfaces* in Figure 4 for an example) can mostly be considered emblems because their gestures replace parts of human speech. Manipulation-based gestures (also see 2.2.2), however, cannot be categorized generally since they can easily be either emblems or illustrators.

One could argue that manipulation-based gestures are illustrators since they are used in the same context as deictic or spatial movements in human-human communication; users simply drop the vocal component because digital systems usually ignore it anyways. In this interpretation, the user's mental model is still focused on the referred object of the interaction and not on the gesture itself (for example, *Charade* in Figure 4). Conversely, one could argue for manipulation-based gestures being emblems because they act as replacements for verbal commands to the digital system. In this interpretation, the pointing gesture substitutes a spoken command not only as an action but also within the mental model of the user. For *Virtual Shelves* (Figure 6), for example, the referred object is abstract or there might not be a referred object at all. The meaning of the gesture then comes exclusively from the gesture itself.

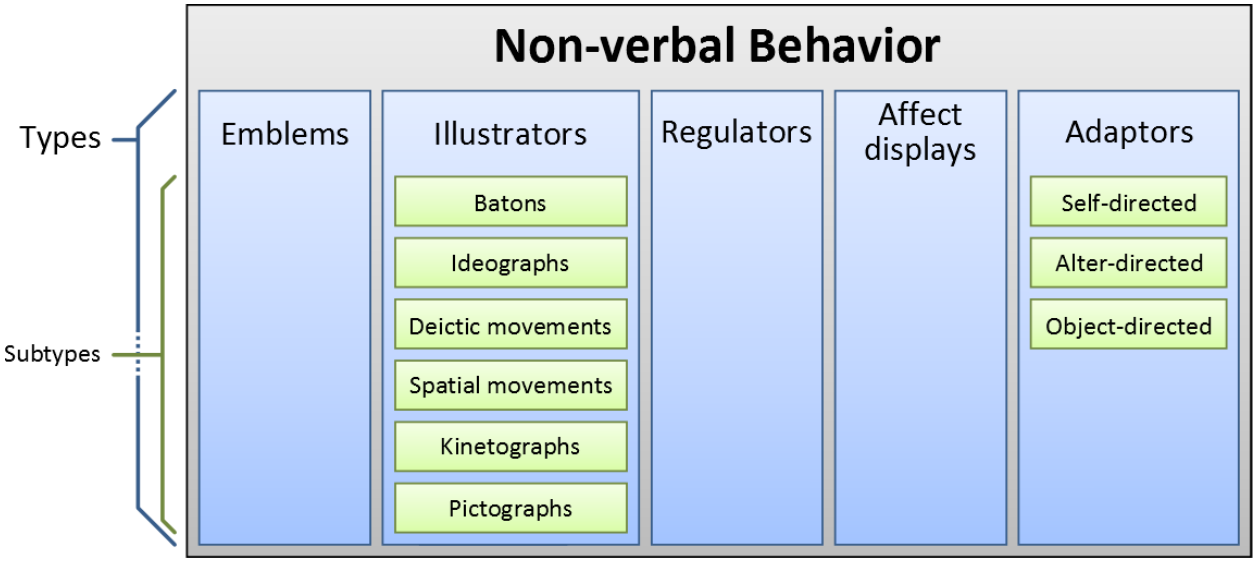



**Figure 11: Categories of non-verbal behavior (adapted from Ekman and Friesen, 1981, p. 102)**

While this distinction might seem to be of minor importance, there is evidence that emblems and illustrators manifest themselves differently in people's mental model. Deictic pointing gestures receive meaning (see 2.5.5 for a rigorous definition of meaning) almost exclusively from the referred object. Emblems, in contrast, receive some meaning from within themselves because of their deep cultural, iconic, and historic background (McNeill, 1992). In the section on human associative memory (2.5.5), I show that people's mental model can have a tremendous influence

on factors such as learnability, memorability, and performance. Having that said, a deeper analysis a comparison of gesture alphabets is outside the scope of this work.

### 2.3.2 Distal Pointing

One way to classify manipulation-based full-arm pointing gestures is by distance between the object and the hand. Researchers differentiate between touch, proximal, and distal pointing (from Latin *distare*: to be far). While the definition of touch is relatively uncontroversial (touch occurs when somatic receptors are triggered), the distinction between proximal and distal pointing is less clear.

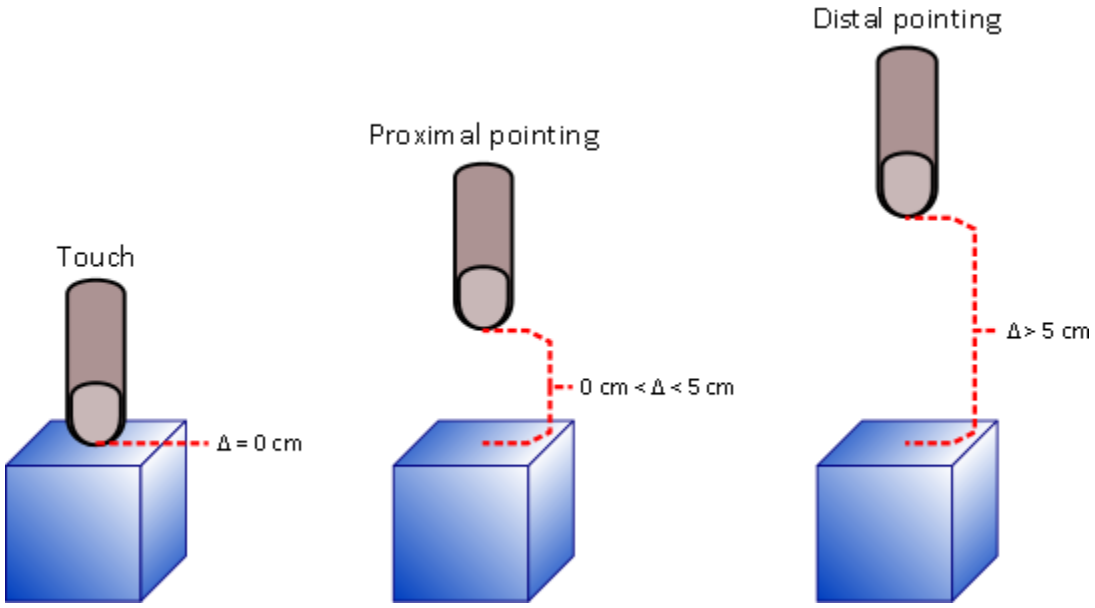



**Figure 12: Hand-to-object pointing distances (values adapted from Povinelli et al., 1997, p. 426)**

A majority of researchers set the threshold between these two types of pointing between 5 cm and 10 cm (Povinelli, Reaux, Bierschwale, Allain, and Simon, 1997). In the context of this work, the exact value is of less importance: when using manipulation-based full-arm pointing gestures in a selection technique, the distance between user and pointing target is usually measures in the meters rather in the centimeters. The pointing gestures I am concerned with can therefore be clearly classified as distal.

## 2.4 Human Sensory, Processing, and Motor Systems

After describing the different types human pointing gestures and their role in human-human communication, I lay out how the production of pointing gestures works on a cognitive level.

First, I introduce the human sensorimotor system and its three main components (2.4.1). Then I describe the role of these three components for the production of pointing gestures (2.4.2 – 2.4.4). Last, I present a method for analyzing the cognitive processes and performance of people's actions (2.4.5).

### 2.4.1 Human Sensorimotor System

Pointing gestures are a small subset of the motions that the human sensorimotor system can perform. The sensorimotor control loop is the same for all human motion (Biedert, 2000).

Humans perceive the environment through somatosensory (tactile, proprioceptive, thermoreceptive, and pain sensation), visual, and vestibular (equilibrioceptive) receptors. This information travels along afferent pathways to the central nervous system, where it is processed in the spinal cord, the lower brain, and the cerebral cortex. From there, commands are issued back to the muscles along the efferent pathways. The execution of these commands by the muscles causes a change in the environment, which is again perceived by somatosensory, visual, and vestibular receptors, thus closing the sensorimotor loop (Biedert, 2000). In the context of human pointing gestures and distal pointing, the proprioceptive and the visual systems are of particular interest, since they play the most important role in the creation of pointing gestures.

The reason why people need this loop to interact with objects in the environment is because people's "perceptual-motor coordination relies upon localizing objects accurately [and] perception is a platform for actions [that] take place in a three-dimensional environment." (Wade and Swanston, 1991, p. 96).

Whenever I refer back to the human sensorimotor system throughout this dissertation, I will color code components of the three sub-systems for better understanding: orange for components of the sensory system, green for components of the processing system, and blue for components of the motor system.

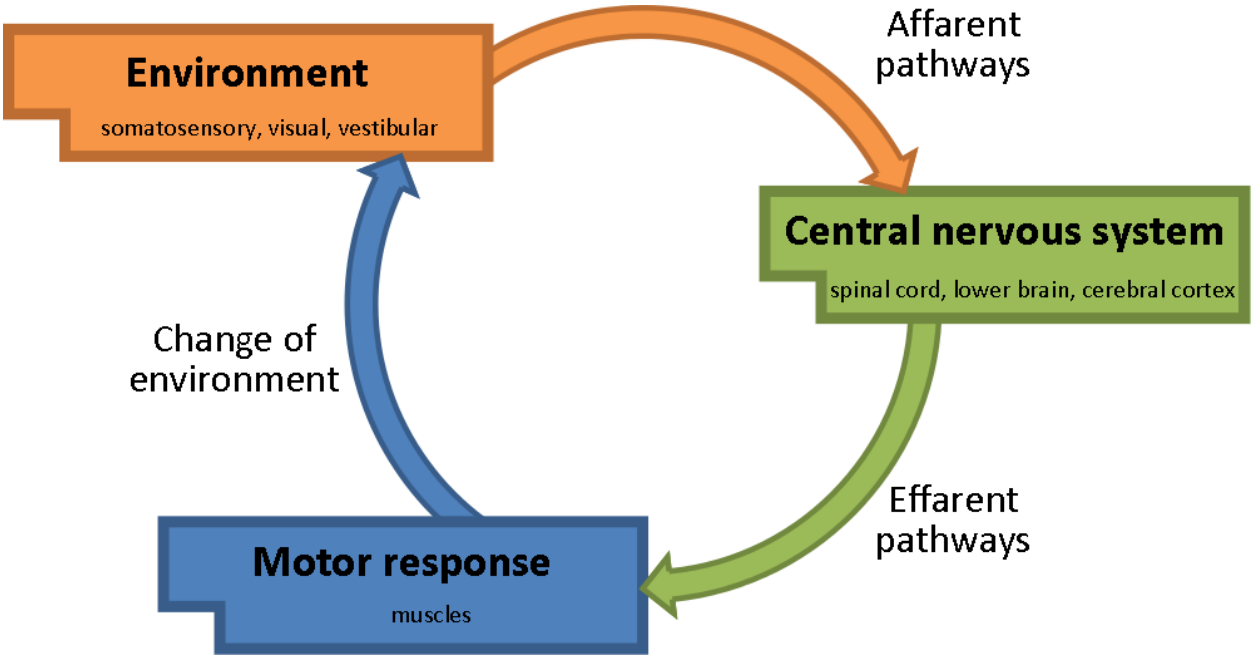



**Figure 13: Sensorimotor system (adapted from Biedert, 1991, p. 23)**

### 2.4.2 Sensory System

#### *Proprioception*

A simple definition of proprioception (from Latin *proprius* ("own") + *captare* ("receive")) is "the sense of body position (conscious and unconscious)." (Hendelman, 2005, p. 250) The term originated from Sherrington in 1906, who stated that the "proprio-ceptive field" contains receptors for several physical stimuli including weight, mechanical inertia, pressures, and strains (Sherrington, 1906, p. 336). This early and relatively broad definition has been challenged in the last few years, and researchers have developed a more detailed approach to proprioception. Lephart et al. give a more specific definition by stating that "proprioception is the acquisition of stimuli by peripheral receptors, as well as the conversion of these mechanical stimuli to a neural signal […]." (Lephart, Rieman, and Fu, 2000, pp. xviii-xix). Distinguished from proprioception are joint position sense, "the submodality of proprioception sense associated with the sense of joint position", (*Ibid.*, p. xxii), and kinesthesia, "the submodality of proprioception sense associated with the sensation of joint movement, either from internal forces (active) or external forces (passive)" (*Ibid.*, p. xxii).

Proprioception plays a key role in most human motor activities and skills. A complete review of this topic is beyond the scope of this work; readers may refer to Lephart et al. for further details (Lephart, Rieman, and Fu, 2000).

### *Vision*

Human visual cognition has been studied for over 2,000 years (Wade and Swanston, 1991). This might not come as a surprise since it is one of the most important systems with which people perceive the world, and since it is the foundation for most human-human interaction. Visual cognition consists of three sub-systems: the visual system, a part of the central nervous system that captures visual stimuli; visual perception and object recognition, a part of the human brain that processes and abstracts visual stimuli into visual imagery; and visual memory, a part of the human memory that stores visual imagery (Luck and Hollingworth, 2008).

In the visual system, rods and cones on the retina of the human eye capture incidental visible light (Wade and Swanston, 1991, pp. 128–134). The human visual field has a size of approximately $\pm 60°$ vertical and $\pm 104°$ horizontal, with an binocular overlap of roughly $120°$ (*Ibid.*, p. 46); the acuity of the average human eye is approximately $0.017°$ (*Ibid.*, p. 53).

### *Proprioception and Vision in Distal Pointing*

Proprioception and vision are the two most important sensory systems involved in executing pointing gestures (Lephart, Rieman, and Fu, 2000). Vision plays a crucial role for creating a spatial understanding of an environment, determining the location of oneself and of the pointing target, and calculating spatial relations between these two. Conti and Beaubaton investigated the effects of occluding the pointing target and the arm movement during distal pointing (Conti and Beaubaton, 1980). They found that lacking either types of visibility impacts people's pointing accuracy, and that observing one's arm movement is more important than observing the pointing target. Biguer, Prablanc, and Jeannerod show that the pointing error caused by lack of visual feedback from the arm correlates with the location of the pointing target relative to the actor: it increased from $3.5°$ for targets $10°$ off the actors viewing direction to over $6.0°$ for targets $40°$ off (Biguer, Prablanc, and Jeannerod, 1984, p. 466). Pointing accuracy toward invisible target increases when people have a better spatial understanding about the environment in which the target is located (Lehnung, Leplow, Haaland, Mehdorn, and Ferstl, 2003). The effect of lack of proprioception is naturally more difficult to study as proprioception is an internal or

interoceptive sense and is thus cannot be willfully ignored. Gordon, Ghilardi, and Ghez compared reaching behavior between neurologically healthy and proprioceptively deafferentiated people (Gordon, Ghilardi, and Ghez, 1995). The authors showed that reaching with neither visual nor proprioceptive feedback produced higher error reaching without vision alone. They also found that vision of the arm at the onset of the motion increases reaching accuracy for deafferentiated people but not for the healthy ones.

### 2.4.3 Processing System

#### *Processing Visual Information*

After the eye has captured image information, it is forwarded through visual sensory memory (or iconic memory) to the working memory (or visual short-term memory (STM)). Iconic memory can be interpreted as simple buffer storage for sensory data: it stores the stimuli coming from the rods and cones in the eye for future processing. Visual STM has four major functions: people use visual STM to combine inputs between saccades from the iconic memory (Luck and Hollingworth, 2008, pp. 73, 76), they use it to detect changes in the environment as an important step for tracking moving objects (*Ibid.*, p. 77), they use it to build up visual long-term memory (LTM) (*Ibid.*, p. 77), and they use it as "a limited capacity system allowing […] such complex tasks as comprehension, learning and reasoning" (Baddeley, 2000, p. 418).

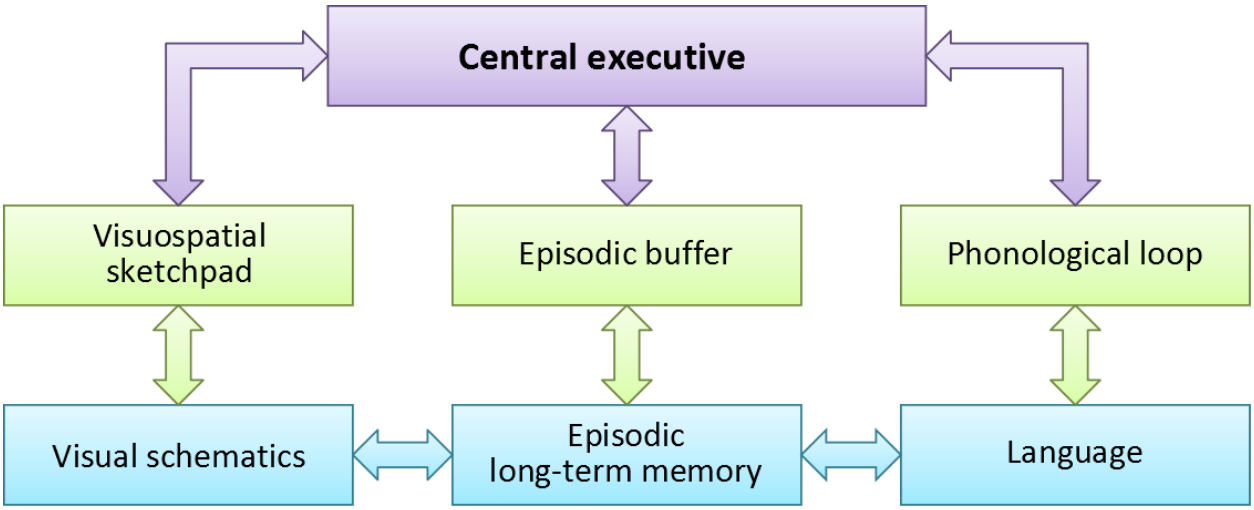



**Figure 14: The current version of the multi-component working memory model (adapted from Baddeley, 2000 , p. 421)**

Visual STM is equivalent with the visuospatial sketchpad in Baddeley's model of working memory (Baddeley, 2000). The visual STM converts visual stimuli into visual imagery through visual perception and object recognition. Visual imagery is then ready to be stored in visual LTM.

Humans use visual LTM to store visual imagery of already encountered objects and environments. Visual LTM "has a remarkably large storage capacity and highly robust retention. […] Visual long-term memory plays a central role in memory for the visual features of objects in the service of object and scene recognition" (Logan, 1988, p. 7). Since vision is the primary sense for people to interact with their world, visual LTM is essential for their daily life and ultimately for their survival, especially since other types of memory depend on the input from visual long-term memory (*Ibid.*, p. 7). One particular feature of visual LTM is its ability to retain scene detail (*Ibid.*, p. 105). Furthermore, people build up scene details automatically and easily as "visual representations are generated and stored in LTM as a natural consequence of viewing" (*Ibid.*, p. 108).

#### *Planning of Distal Pointing Gestures*

The process of planning distal pointing gestures and the differences in the planning process depends on the pointing target are core aspects in the analysis of room-based interaction. Much of this process can only be understood with firm knowledge about procedural memory, the part of human memory that governs the planning and execution of human movements. I cover procedural memory and skill acquisition and execution in section 2.5.4.

### 2.4.4 Motor System

#### *Kinematics of Distal Pointing Gestures*

Kinematics is the theory of the motion of bodies, and deals with mathematical descriptions of motion (Benenson, Harris, Stöcker, and Lutz, p. 3). Soechting and Lacquaniti investigated the kinematics behind distal pointing (Soechting and Lacquaniti, 1981). One of their findings was that "the trajectory in space described by the movement differs little from trial to trial and is independent of the speed of the movement" (*Ibid.*, p. 718). This implies that once people have learnt to perform a pointing gesture at a target, they can reproduce it at different velocities without having to relearn the necessary kinematics. Furthermore, the authors showed that

shoulder and elbow movement are tightly coupled, whereas the wrist moves independently. While this has no major influence on the execution of distal pointing, it expresses the viability of solely using ones wrist to perform pointing gestures.

Biguer et al. investigate the accuracy with which people can point at horizontal targets between $-40°$ and $+40°$ with coordinated (head can move freely) and uncoordinated (head is locked) pointing (Biguer, Prablanc, and Jeannerod, 1984). They found that the absolute error did not exceed 2.5° as long as the target was in people's foveal vision. When the pointing target disappeared into peripheral vision, accuracy quickly became worse.

Every gesture consists of a series of up to six phases (McNeill, 2005). Out of these six phases, the stroke is the only obligatory phase, as in the "absence [of] a stroke, a gesture is not said to occur" (*Ibid.*, p. 32). The stroke and stroke hold are also "the [only] gesture phase with meaning" (*Ibid.*, p. 32). The difference between stroke and stroke hold is that the actor's body movies during the first but remains static during the latter. The purpose of the preparation phase is setting up the body in a position to start the gesture stroke. Pre- and post-stroke holds are there to separate the stroke form the preparation and retraction phases and thus emphasize the stroke. In the retraction the actor's body moves back into a neutral content- and effort-free state.

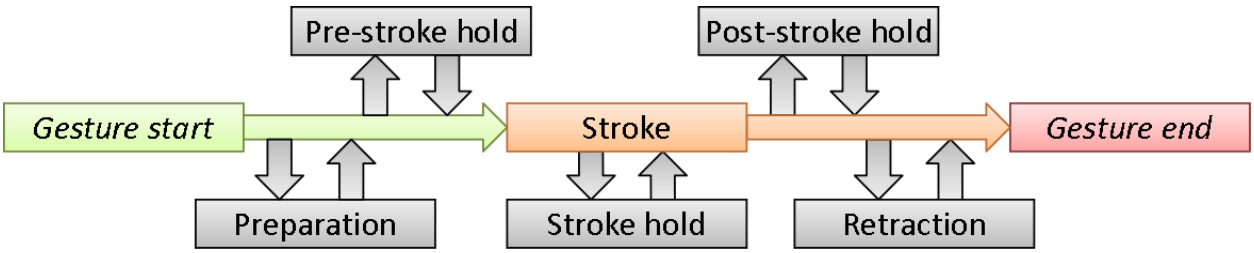



**Figure 15: Anatomy of a human gesture; gray elements are optional (adapted from McNeill, 2005, pp. 31–33)**

### *Interpretability of Distal Pointing Gestures*

Interpretability of a pointing gesture expresses how well a person or a digital system can infer the actual pointing target and the underlying meaning from the gesture. There are multiple ways to approach this problem. One example is to restrict oneself to the purely physical aspects of the gesture, such as origin and direction. Bangerter and Oppenheimer used this approach and asked

participants to specify at which marker on a field of equally-spaced markers the experimenter was pointing (Bangerter and Oppenheimer, 2006). The authors reported an average pointing error of 3.5° for horizontally-arranged targets and 2.5° for vertically-arranged targets. In a combined task, the errors dropped to 2.5° (horizontal) and 1.8° (vertical).

While the physical aspects of pointing gestures are important for assessing the target of a distal pointing gesture, they are not the only ones. Especially in human-human communication, the context in which a gesture is performed helps observers to interpret its target and meaning. There are three major theories for mechanisms that provide this context: constraints (Nelson, 1988), shared attention mechanisms (Baron-Cohen, 1995), and intentions (Tomasello, 1995). While the exact mechanisms are still debated, their effects are not doubted. Schmidt investigated these effects by showing video-taped distal pointing gestures from infant-parent interaction to participants. He observed that "attention-directing gestures are not ambiguous for observers and [the results] support the view that gestures are intrinsically related to what they indicate." (Schmidt, 1995, p. 161)

### 2.4.5 The GOMS-Model and the Model Human Processor

There exist several approaches for a theoretical analysis of human cognitive performance under certain tasks. Card et al. introduced two, now widely accepted, complementary methods for dissecting the cognitive units of a task: the GOMS-model and the Model Human Processor (MHP) (Card et al., 1983).

The GOMS-model (short for goal–operator–methods–selection) splits high-level tasks into smaller units. The individual units can then be analyzed in terms of, for example, completion time or cognitive load. In the GOMS model, the *goal* is the overall state that a user wants to achieve. "A *method* describes a procedure for accomplishing a goal. It is one of the ways in which a user stores his knowledge of a task" (Card, et al, 1983, p. 145). Occasionally, there is more than one suitable method for achieving the goal. The selection rule is a metric for determining which method to use. When these selections occur smoothly and quickly, people achieve skilled behavior. Skilled behavior is closely related to the aforementioned idea of routines as glue of everyday life (see 2.1.2). Finally, "operators are elementary perceptual, motor, or cognitive acts, whose execution is necessary to change any aspect of the user's mental

state or to affect the task environment" (*Ibid.*, p. 144). Each method normally consists on several operators that have to be completed in order to achieve the overall goal.

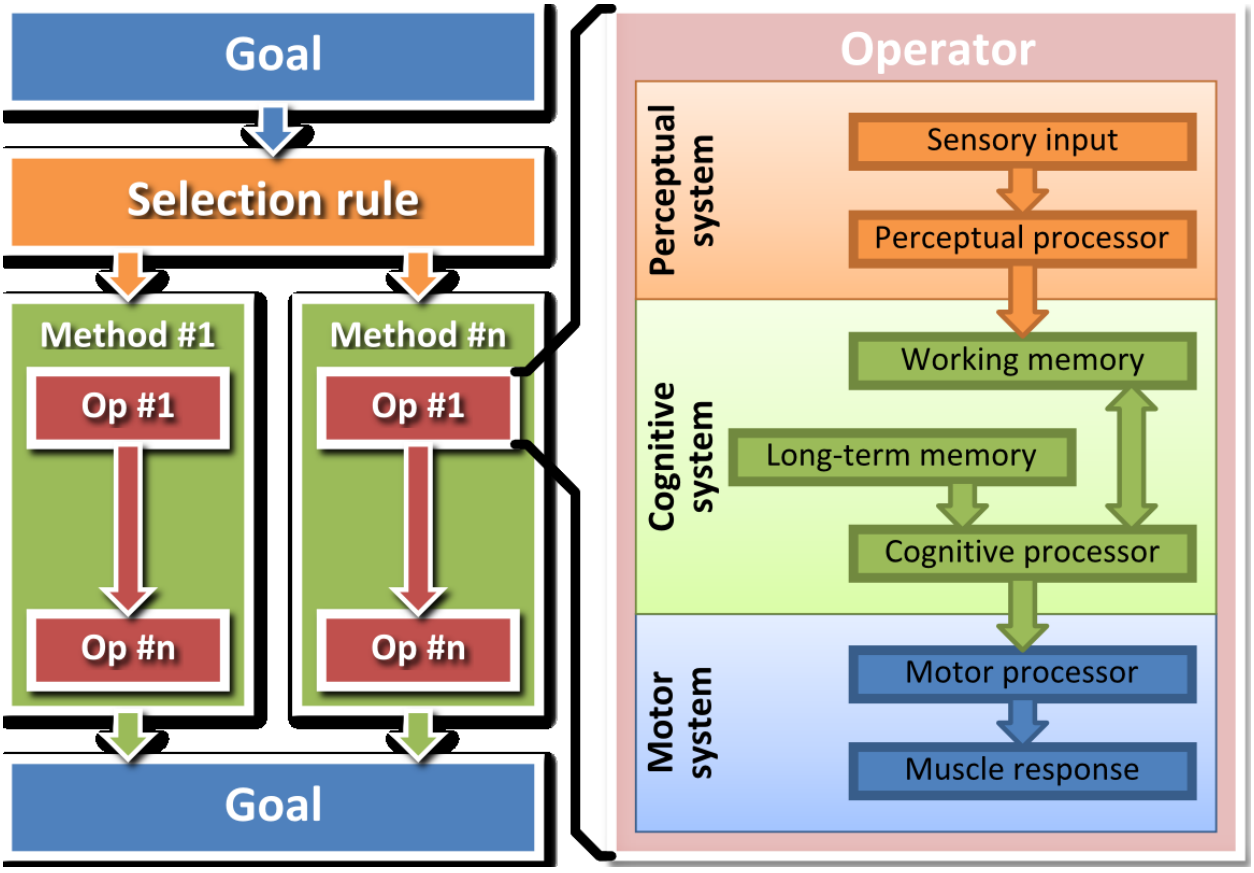



**Figure 16: GOMS-model (left) and MHP (right); (adapted from Card et al, 1983, pp. 26, 144–145)**

The model human processor (MHP) complements the GOMS model in the sense that it can be used to analyze a single previously identified operator. For this, MHP breaks down an operator in three phases: the perceptual phase, in which users capture and pre-process sensory input; the cognitive phase, in which the human brain devises an appropriate strategy for completing the operator; and the motor phase, in which the human brain calculates the necessary (efferent) neural signals for controlling the correct muscle groups (Card et al., 1983). As described above, the validity of the MHP has been confirmed by other researchers, for example, in Biedert's model of the sensorimotor system (see Figure 13) (Biedert, 1991). The interaction between the working memory and the cognitive processor also resembles the interaction between buffers,

sketchpads and the central executive in Baddeley's model of the working memory (see Figure 14) (Baddeley, 2000). I described the perceptual system and the motor systems relevant for performing pointing gestures earlier in this chapter (see 2.4.1 and 2.4.2), and I will introduce all relevant cognitive systems in the section on human memory systems (see 2.5).

While the main purpose of both the GOMS-model and the MHP is (quantitatively) estimating people's task completion time, they can both be used for a (qualitative) comparison of interaction techniques, in which researchers assess the structure of an interaction without assigning concrete performance values to each cognitive unit.

## 2.5 Human Memory System

In this section, I examine the memory systems relevant to room-based interaction, i.e. learning semantic information and creating pointing gestures. I start with a short definition (2.5.1) and a taxonomy of human memory (2.5.2). I then will talk about the three memory systems relevant to room-based interaction and other types of selection techniques that I will address in my dissertation: spatial memory (2.5.3), procedural memory (2.5.4), and semantic memory (2.5.5).

Overall, this chapter will reflect a high-level cognitive approach to human memory. Other approaches, for example from neuropsychology, are also relevant to human pointing gestures, but they operate on a lower level and therefore contribute less to a general understanding of the subject; an analysis of these approaches is outside the scope of my dissertation.

### 2.5.1 Definition of Human Memory

Human memory is a system of processes that describe and conceptualize how information is encoded, stored, and retrieved (Dudai, Roediger, and Tulvin, 2007, p. 11). Among psychologists, learning is oftentimes considered to be synonymous to memory because both terms describe the same underlying effect: experience-dependent behavior (Elias and Saucier, 2006, p. 207). As the term experience-dependent behavior indicates, learning can be defined as a relatively permanent change in human behavior as a result of some past experience (*Ibid.*). Similarly, memory is a record of past experience that causes a change in peoples' behavioral or cognitive capabilities (Anderson and Bower, 1980, p. 42).

### 2.5.2 Taxonomies of Human Memory

Although there are numerous ways of categorizing human memory, two major approaches dominate. The first one—a bottom-up approach—is rooted in neuropsychology and tries to categorize human memory and its subsystems by functional groups within the human brain, such as lobes, cortices, and pathways (Squire and Zola, 1996, p. 13516). The second one—a top-down approach—is rooted in cognitive psychology and behaviorism and tries to categorize human memory and its subsystems by people's behavior and performance on certain tasks. Generally, the first approach only allows for a narrow and constrained model of human memory while validating itself with strong evidence from clinical research. The second approach, on the other hand, provides us with a richer, more detailed model of human memory while sometimes being more speculative and unstructured.

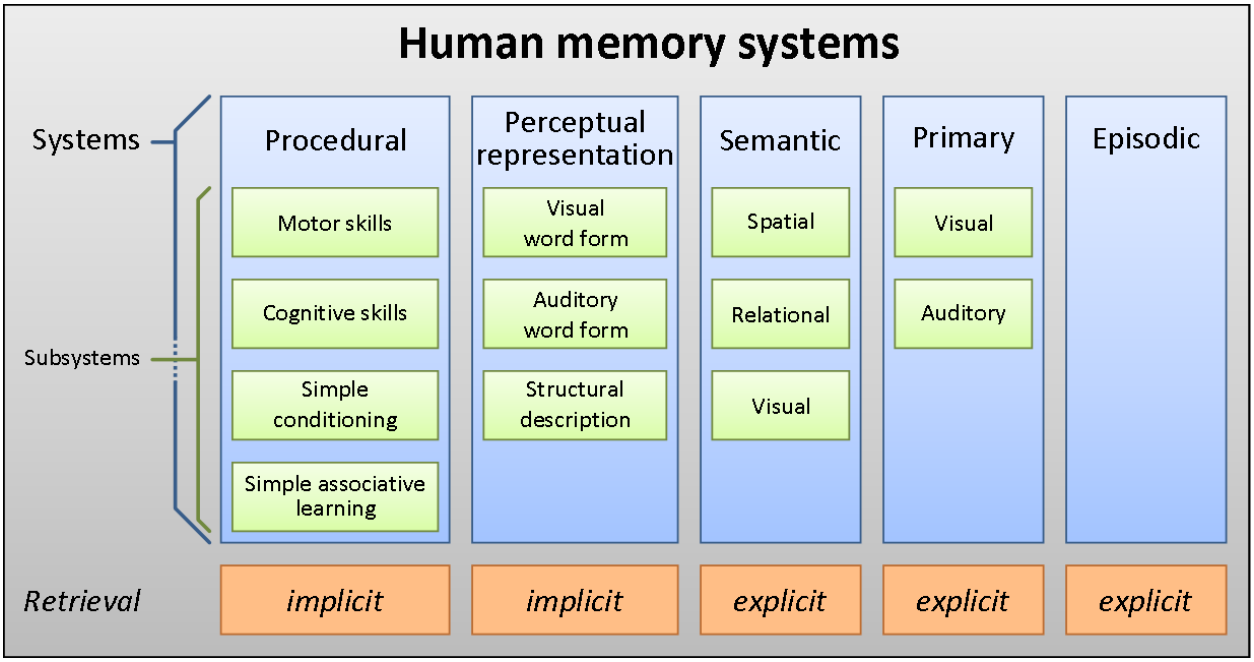



**Figure 17: Major systems of human memory (adapted from Schacter and Tulving, 1994, p. 26)**

#### *Cognitive Approach to Human Memory*

The currently accepted version of a cognitive memory taxonomy was published by Schacter and Tulving (Schacter and Tulving, 1994, p. 96). This taxonomy was developed from the idea of cognitive maps by Tolman who was the first to present evidence for the existence of multiple

forms of memory (Tolman, 1948). Over the years, this taxonomy evolved into the five-system view shown below. This particular division is supported by results from numerous experiments with participants who suffered from different neurological disorders, such as amnesia, agnosia, or apraxia caused by conditions like dementia, stroke, or epilepsy. This division reflects the current state of research but is by no means final: researchers occasionally argue for adding new memory systems, such as emotional memory (Nalbantian, 2011, pp. 277–296) or new memory subsystems, such as visual (long-term) memory (Palmeri and Tarr, 2008, pp. 163–208) (this new subsystem should not be confused with the visual subsystem of primary memory, which is short-term; see 2.4.1).

### *Other Commonly Used Approaches*

Two other approaches are oftentimes used, especially outside the field of psychology. Although they are hampered by their over-simplification of memory processes, I describe them briefly due to their frequent use in literature.

#### Declarative—Non-declarative

A simple and comprehensible distinction between declarative and non-declarative memory is that declarative memory contains all factual accumulated knowledge, e.g., that Paris is the capital of France or how one's grandmother looks like, whereas non-declarative memory contains all information about how to perform a certain action, e.g., how to ride a bicycle or how to behave after conditioning (Cohen and Squire, 1980, p. 209). An important property of non-declarative memory is that "knowledge represented in this system is not consciously known and cannot be transferred from one person to another" (Surprenant and Neath, 2009, p. 11). Nevertheless, many researchers agree that the distinction between declarative and non-declarative memory is too coarse a concept for understanding the nature of multiple memory systems (Schacter and Tulving, 1994, p. 51).

#### Implicit—Explicit

The distinction between implicit and explicit memory is similar to the one between declarative memory and non-declarative memory (Schacter and Tulving, 1994, p. 233), yet subtle differences remain and are still debated (Roediger, 1990, p. 374). One common definition is that information from explicit memory has to be recollected in an intentional or conscious act, whereas information from implicit memory does not require such act (Schacter and Tulving,

1994, p. 233). However, implicit and explicit memory do not refer to particular systems in the human brain but are rather a way for distinguishing between the behavior of different memory systems (Roediger, 1990, p.373).

In the next section, I give an overview of spatial memory and motor skills, two memory subsystems that are necessary for identifying, finding, and interacting with real-world objects.

### 2.5.3 Spatial Memory

The first component of human memory that is relevant to this research is spatial memory. Spatial memory is defined as "a record of geometric relations involving observers, objects, and surfaces" (Allen, 2003, p. 42). Spatial memory plays an essential role in peoples' daily life; they use it to navigate the world, find objects within environments, and interact with items. Therefore, it is reasonable to assume that people generally have a well-developed spatial memory. It is important to realize that space in the context of human spatial memory "is a multifaceted construct that includes both real space and imagined space" (Elias and Saucier, 2006, p. 323).

#### *Learning, Remembering, and Performance*

It is well established that spatial locations can be remembered with great accuracy and without great effort (Allen, 2003, p. 44). People are able to subconsciously build up a spatial model of their environment; in fact, "it is generally agreed on that observers update the representation of the target's location as they are locomoting" (*Ibid.*, p. 165). Building up spatial memory is often a byproduct of locomotion, one of a human's core abilities. While this process is by no means effortless, it does not require much attention and it happens automatically (that is, people cannot choose to not acquire spatial memory).

There are three ways that humans encode spatial information: in categories, in coordinates, and through perception-action. Categorical coding is "a robust means of remembering spatial information based on the gestalt of the environment", whereas coordinate coding "involves conceiving of objects or events existing in an abstract space consisting of an infinite number of points organized by a coordinate system" (Allen, 2003, p. 59). This distinction is important because it explains how humans perceive their environment: people remember the location of real-world objects either by their spatial relationship ("the book is in the book shelf right next to the TV") or by their location ("the book is about 10 cm left of my left hand"). From these two

examples it becomes obvious that coordinate coding requires some sort of coordinate system—mostly Cartesian or spherical—and an origin—mostly self-centric.

Categorical coding of spatial information is a highly efficient process in the sense that it creates relatively accurate results for very little effort (Allen, 2003, p. 59); people create it is mostly automatically during locomotion (*Ibid.*, p. 16). Coordinate coding, in contrast, requires higher cognitive effort but yields more accurate results than categorical coding (*Ibid.*, p. 60). From this it becomes apparent that people use categorical and coordinate coding for different purposes.

People use categorical coding to gain a rough understanding about a given environment: what objects are present?, where are they located?, what might be their function?. In addition, people also spatially relate present objects to each other and thus build a hierarchical map of objects in an environment. The level of detail to which people discern objects depends on several factors, including character (Gestalt) of the object or environment, (visual) perception, and current activity (Allen, 2003, p. 59). As mentioned before, people create a categorical model of an environment subconsciously and without much mental effort.

Coordinate coding, in contrast, is useful when people require spatial information with high precision. The most common use of coordinate coding is when people want to physically interact with an object or want to avoid interaction. In this case, people have to precisely know the location of an object so that they can initiate appropriate motor or locomotive actions. Again, people cannot create a coordinate model of an environment subconsciously; the process requires explicit mental attention and effort.

The last possible way of coding spatial information, the perception-action system, refers to spatial knowledge that people gather as a direct result of their motoric actions (Allen, 2003, p. 43). People automatically gather spatial information about objects or the spatial relation between objects by, for example, walking through an environment: they collect information about the height of stair treads and the length of a hallway. This sort of spatial information, however, decays within seconds and is therefore not relevant in the context of this research (*Ibid.*, p. 49).

The first step of remembering object location is conjuring a visual or abstract representation of the object (Allen, 2003, p. 145). Second, the previously stored binding between object and

location is retrieved from spatial memory. Last, categorical and coordinate information are decoded.

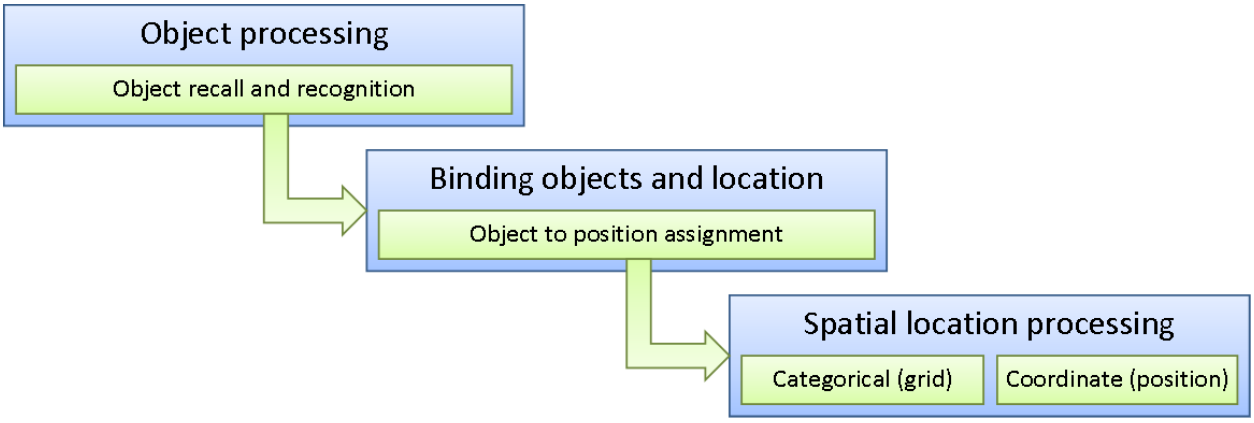



**Figure 18: A functional analysis of object-location memory (adapted from Allen, 2003, p. 145)**

In the context of this research, it is difficult to give generalizable performance measures, such as accuracy, cognitive load, and time requirements, for human spatial memory (*Ibid.*, p. 59).

### *Coordinate Systems and Reference Frames*

As mentioned above, people have to use coordinate systems to organize their spatial memory (Allen, 2003, p. 5). People use basic elements of an environment for creating the coordinate system. On the most basic level, gravity, determines what axis people perceive as "up—down" (z-axis); the shape of a room or the arrangement of objects determines if people use Cartesian or spherical coordinate systems; the edges of a room determine the orientation of the floor (x- and y-axes) (*Ibid.*, p. 5). People use basic patterns and shapes, similar to the one described by Gestalt theorists, such as circles, triangles, or rectangles, to set up coordinate systems and relationships between objects (*Ibid.*, p. 59).

Reference frames are not only important for coordinate coding but for categorical coding as well. People use the same basic patterns and shapes to define spatial relationships between objects. For example, vertically arranged objects oftentimes indicate a hierarchy where the "higher" stand for "more" or "first" and lower stands for "less" or "later"; horizontal lines indicate equality between objects; rectangles indicate some sort of grouping (Gillie and Broadbent, 1989).

### *Creating and Updating Spatial Models*

Creating a spatial model of an environment is divided into three steps. "First, the various items in the to-be-remembered display need to be processed. […] Second, a component might be distinguished that is relevant for processing the necessary location information. […] Finally, common to all visual processing and central in object-location memory, the object-identity information and the spatial information need to be combined" (Allen, 2003, pp. 144-146).

The first step refers to the visual task of capturing present objects in the room. Which object a person actually captures depends mostly on his visual capabilities, the regions of the environment he is scanning, and the visibility of the objects. The second step applies to the objects that people have visually captured. In this step, the person identifies and labels the captured objects ("this is the mug that…") and processes their spatial locations. Whether captured objects are chosen for identification and with what levels of detail the identified objects are labeled depend on the person's task and his familiarity with the objects. Furthermore, the person generates and processes spatial information about recognized and labeled objects, particularly their categorical location ("on the desk, next to…") and their coordinate location ("half my arm's reach in front of me"). In the final step, identity and spatial information are merged into one memory element. It is currently assumed that "position processing might be mostly automatic, whereas identity processing requires central effort" (Allen, 2003, p. 146).

### *Spatial Memory and Full-arm Pointing Interaction*

Distal pointing is most accurate in a closed-loop feedback condition (see below and 2.4.2), whereas a person's distal pointing performance suffers in uncoordinated feedback systems. This negative effect, however, becomes weaker when people are in familiar environments (Lehnung, Leplow, Haaland, Mehdorn, and Ferstl, 2003). A detailed spatial model of an environment can therefore improve people's distal pointing performance.

When a digital system uses real-world proxies for interaction, people have to find the proxy-object that is associated with the desired interaction. Keeping a proxy-object static within the environment increases the speed with which its location is stored in spatial memory. The proxy-object is now part of a person's spatial model of the environment; this reduces the time required to find the proxy-object.

### 2.5.4 Procedural Memory and Motor Skill

The second main memory system relevant for this research is motor skill, which is a subsection of procedural memory. "Procedural memory enables organisms to retain learned connections between stimuli and responses, including those involving complex stimulus patterns and response chains, and to respond adaptively to the environment" (Tulving, 1985, p. 387). This definition includes a wide variety of different tasks, similar to the model of Schacter and Tulving (Schacter and Tulving, 1994). Information stored in procedural memory has no absolute values and is internal: that is, it contains information about the person and not the environment. Procedural memory shows implicit behavior as it operates at an automatic rather than consciously controlled level. Lastly, procedural memory cannot be directly transferred between people (*Ibid.*, p. 26).

A skill is a defined sequential or hierarchical set of activities; using a skill means executing the required set of activities (Fitts and Posner, 1967, p. 1). An activity is either an innate human function or a previously acquired skill (*Ibid.*, p. 3). With this definition, one can easily see skill acquisition as a process that starts with birth, where infants can only execute innate functions, and continues throughout life: earlier learned skills are the foundation for learning new skills. The execution of a skill usually involves closed-loop feedback from the sensory system (see below and 2.4.2). People subconsciously use this information in order to modify and improve the currently ongoing skill execution. Furthermore, people use this information from previous trials to improve their overall skill level (*Ibid.*, p. 2). Feedback from people's sensory system in combination with the outcome of an executed skill is a useful tool for improving one's skill level (*Ibid.*, p. 12). I discuss this so-called intermediate learning phase later.

Generally, scientists differentiate four classes of learned skills: gross bodily skills, manipulative skills, and perceptual skills, which all involve responses to real objects in the spatial world, and language skills, which involve the manipulation of signs and symbols (Fitts and Posner, 1967, p. 4). For the purpose of this work, I focus on manipulative skills.

#### *Learning, Remembering, and Performance*

From the previous definition of skill as a set of activities, we can easily define learning as improving performance in the proper execution of this set (Fitts and Posner, 1967, p. 8). Overall,

skill learning happens in three stages: the cognitive, the associative, and the autonomous phase (*Ibid.*, pp. 11–15).

In the cognitive stage, students of a skill take a set of existing skills and arrange their sequence to form a new skill. Beginners learn the sequence by observing it from abstract instructions or concrete demonstrations (Fitts and Posner, 1967, p. 11). Beginners also have to memorize and semantically conceptualize this sequence before executing it for the first time (*Ibid.*, p. 11). Students complete this stage once the concept of the sequence is completely memorized. As we will see, this semantic concept of the sequence will fade once students reach a sufficient proficiency in a skill.

During the associative stage, beginners transform a set of activities into a new skill by practicing and improving the transitions between activities. The intermediate phase can take up most of the time of learning a new skill; the length depends on the size of the activity set, i.e. the complexity of the skill, and the individual competence of the student (Fitts and Posner, 1967, p. 12).

In the autonomous stage, beginners stop seeing their actions as a sequence of activities and rather experience it as a single skill (Fitts and Posner, 1967, p. 14). The underlying semantic concept of the sequence is now unnecessary for the execution of the skill and might be forgotten until an expert has to re-conceptualize it because they want to teach the skill to beginners. This also implies that experts do not have to access semantic memory anymore if they want to execute a skill because the entire skill sequence is now stored in procedural memory. Given the nature of procedural memory (see above), experts can now perform a skill automatically without conscious effort, thus reducing their cognitive load drastically. Ultimately, a skill can become close to a reflex (*Ibid.*, p. 15).

A general approach to measure the proficiency of a skill is to measure the accuracy and uniformity of the involved activities (Fitts and Posner, 1967, p. 2). Performance quality of a skill is simply the degree of consistency and precision with which people are performing the skill. Performance of any skill is limited to how much improvement is possible as a result of practice. It is difficult, however, to predict the rate of skill performance increase. Generally, the rate of improvement is reduced as practice continues. This relationship can be described as a power function (*Ibid.*, p. 18). While the idea of a power function being the general shape of the learning

curve has been established through countless experiments, the slope of this curve (i.e. the difficulty of learning) depends of the complexity of the skill. Another approach for determining skill proficiency is measuring fluency, which is the combination of execution speed and accuracy (MacKay, 1982, p. 483). When executing a skill, people can generally chose to perform it with an emphasis on execution speed or accuracy. This effect is called "speed–accuracy trade-off, one of the most reliable and pervasive phenomena in the study of skilled behavior" (*Ibid.*, p. 495).

The exact cognitive processes involved in motor skill acquisition are still unknown. A widely accepted theory is the closed-loop theory (Adams, 1971), which states that humans compare "sensory feedback from the ongoing movement […] with the stored memory of the intended movement" (Shumway-Cook and Woollacott, 2001, p. 31). Although the closed-loop theory can explain many phenomena related to motor skill acquisition, it has come under scrutiny (*Ibid.*, p 31 – 32). Schmidt's schema theory provides a more general approach for explaining the acquisition of procedural memory (Schmidt, 1975). At the core of his theory are motor response schemas, generalized motor programs that humans can modify to create the desired outcome. Schmidt also lists four memory schemas that humans store in procedural memory: the initial movement conditions, the desired outcome of the movement, the parameters of the generalized motor program gathered from previous executions (result knowledge), and the sensory information during the movement (error assessment) (Shumway-Cook and Woollacott, 2001, p. 32). According to the schema theory, the motor program is derived from motor response schemas, which are rough templates of motor activities. These schemas are adapted using the initial conditions of the human body in the environment, the desired outcome of the movement, the knowledge of the outcome of previous executions of similar movements, and a closed feedback loop about the current outcome of the movement. In schema theory, two processes play an important role in learning: result knowledge and error assessment. Result knowledge is the knowledge of how input parameters (e.g., muscle flexion amplitude and movement timing) determine movement output. Error assessment is a mechanism for evaluating and labeling discrepancies between current and desired outcome, and current and expected proprioception and exterioception. Normally, humans use both processes concurrently and in varying ratios, although medical conditions, such as somatosensory deafferentation, can disable error assessment. Using error assessment, however, is less accurate, thus slower and cognitively more demanding, than using result knowledge (Schmidt, 1997). Subsequently, improving motor skills

is simply shifting the ratio between result knowledge and error assessment toward using result knowledge.

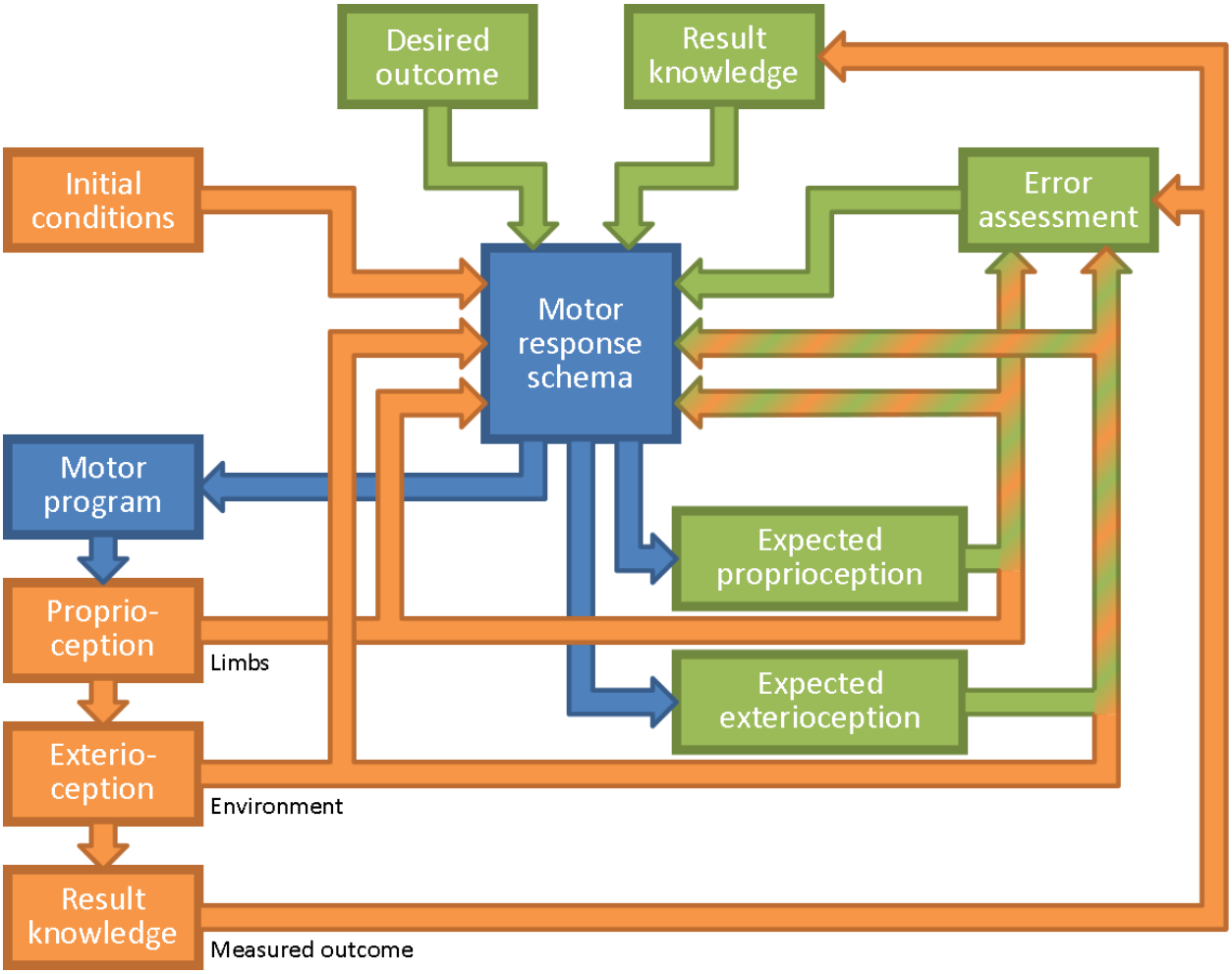



**Figure 19: Schema theory of discrete motor skill learning; orange: perceptual system, green: cognitive system, blue: motor system (adapted from Schmidt, 1975, p. 238)**

When humans perform a motor skill for the first time, they have to rely completely on error assessment as they have not built up any result knowledge. The resulting movement is subsequently inaccurate, and humans have to iterate through the control loop more often, which slows the movement and causes higher cognitive load. After sufficient practice, humans will have built up a substantial amount of result knowledge and they can use this knowledge when modifying the motor response schema in order to calculate the motor program. The initial motor

program is therefore much more accurate and requires less iterations through the control loop. The person has reached the autonomous phase of motor skill acquisition.

### *Transition from Semantic Memory to Motor Skills*

For this research, the transition from the associative phase to the autonomous phase is of major interest. It also marks the transition from the use of semantic memory to the use of procedural memory. Since experts can now perform a skill automatically without conscious effort and with reduced cognitive load, completion of the task that involves this skill becomes less demanding and more accurate. The underlying neurological reason for the reduced conscious cognitive load is that toward the autonomous phase, the human brain bypasses the working memory, i.e. the part of human memory in which conscious decision making occurs (Beilock, Wierenga, and Carr, 2010). Unfortunately, there is almost no research that looks in a generalizable way at the transition from the associative to the autonomous phase (Schmidt and Lee, 2005, p. 431; Shumway-Cook and Woollacott, 2001, p. 38).

Furthermore, the criteria that mark this transition are still debated (Logan, 1988, p. 515). Having that said, it is generally acknowledged that the time needed to make this transition is measured in the magnitude of months or years (see Newell and Rosenbloom, 1981; and Fitts and Posner, 1967, pp. 15–19 for examples).

For the purpose of this work, however, it is useful to present at least one metric to assess people's learning success. One of the most common laws to predict completion time of a task given a certain amount of practice is the power law of practice. It was first mentioned by Newell and Rosenbloom in 1981 (Newell and Rosenbloom, 1981). Newell and Rosenbloom wrote the law as $T(N) = A + B \cdot (N + E)^{-\alpha}$, where $A$ is the minimum completion time, $B$ the time on the first trial, $N$ the number of trials, $E$ the number trials from prior experience, α the slope of the line, and $T(N)$ the time needed to complete the task on the $N$-th trial. One should keep in mind, however, that the "law of practice is just a description of the relationship between practice trials and performance. […] this relationship does not necessarily provide a description of the process of learning—that is, the underlying capability for performance that is the goal of practice and learning research" (Schmidt and Lee, 2005, p. 323). This means that the power law, for example, does not make a statement about the effort that a person has to make in order to perform the task.

#### *Skills and Full-arm Pointing Interaction*

Performing full-arm pointing gestures is one of peoples early acquired general skills. Repeatedly pointing at the same proxy-object, however, can turn from a series of activities into a new skill. If this transition occurs, people would be able to perform device interaction through full-arm pointing gestures with higher accuracy and lower cognitive effort than before.

### 2.5.5 Associationism and Semantic Memory

Associationism is a particular way of conceptualizing processes within the human brain. It interprets memory as a collection of associations between stimuli and responses created by experience. Stimuli and responses can be ideas, sensory data, or memory nodes in the mind (Anderson and Bower, 1980, p. 10). People memorize these associations because they have occurred together in the past. Other than this co-occurrence, there is no restriction on what stimuli and responses are or how they are connected. As I show in the following section, associationism is a useful concept for understanding and predicting human memory performance. This makes associationism directly relevant to learnability and memorability of real-world selection proxies (see 2.2.5), which are important in the context of human pointing gestures—in particular, because users can associate proxies with objects of interest.

As with many psychology-based theories regarding human memory, associationism only provides an abstract framework and does not give concrete answers about the neuropsychological function of the human brain, nor does it root itself in medical research (Anderson and Bower, 1980, p. 70). Instead, associationist theories aim at simulating human memory with computer programs (*Ibid.*, p. 70). As a result, the language in which psychologists describe processes in human associative memory is oftentimes similar to formal languages from the field of programming language theory and linguistics (for an example, see *Ibid.*, p. 69).

Aside from psychologists, computer and software engineers have studied human memory from an algorithmic perspective (Kohonen, 1984, p. 4). An early attempt to algorithmically describing associative memory in form of a neural network was made by Willshaw et al. (Willshaw, Buneman, and Longuet-Higgens, 1969). Their work was the basis for an influential paper on neural networks , published by Hopfield (Hopfield, 1982).

The psychological and the algorithmic approaches focus on different aspects of associative memory. Researchers of neural networks have the goal of building computational systems that are modeled after the brain's neurons and synapses (Hopfield, 1982, p. 2556), whereas associationists are more concerned with describing and evaluating the performance of human associative memory. For the purpose of this work, I focus on the associationist view.

### *Learning, Remembering, and Performance*

Learning in associationist theories—often referred to as paired associative learning—is defined as "associating the cue term appropriately to the response term […] This is always done by propositionalizing the relationship—either finding a preexisting relationship between the two concepts corresponding to the stimulus and response term, or confabulating an 'artificial' relationship to deal with the exigencies of the learning task itself" (Allen, 2003, p. 189). In this context, a proposition is a "configuration of elements which […] conveys an assertion about the world" (*Ibid.*, p. 3).

From this definition, one can draw several conclusions about the function and the capabilities of associative memory. First, associations between stimuli and responses depend on previous experience. Therefore, a group of people can share the same associations if they have shared the same experience. It also implies that certain associations are more commonly known (e.g., $red \rightarrow warm$) than others (e.g., $\text{⌘}Q \rightarrow quit\ program$). Second, stimulus and response do not necessarily have to be "obviously" related; instead, people can create any—for outsiders probably obscure—relationship between stimulus and response. These confabulated relationships are less likely to be commonly known since they often times refer to a past experience of the particular person and are otherwise unrelated to the current stimulus or response. Third, associationism does not restrict the nature of either stimulus or response. Stimuli and responses can be verbal, imagery, or any sort of sensory input; a response itself can even be a stimulus for another response, an effect called associative chaining (Johnson, 1969). Chaining is one of the two major methods for reducing memory load (the other one is chunking). Chaining means that a response becomes a stimulus $S$ for the next response $R_i$: $S \rightarrow R_1, R_1 \rightarrow R_2, R_2 \rightarrow R_3, etc$. With chunking, in contrast, all responses are subsumed in one chunk $C$: $S \rightarrow C \rightarrow R_1, S \rightarrow C \rightarrow R_2, S \rightarrow C \rightarrow R_3, etc$ (see *Ibid.* for a comparison).

As stated above, a proposition is a "configuration of elements which […] conveys an assertion about the world" (Anderson and Bower, 1980, p. 3). According to associationist theories, propositions are represented in deep structure tree-diagrams, which originate from Chomsky's representation of language grammar (*Ibid.*, p. 81; for examples, see *Ibid.*, p. 85, and Chomsky, 1965, p. 65). While this aspect is not of major relevance here, it will become important later on when discussing the structure of multiple complex associations, such as associative chaining.

Remembering in associative memory is the process of retrieving a response term given a certain cue term. An interesting question here is whether the relationships between stimuli and responses are uni- or bi-directional. Uni-directional (following the independent association hypothesis) implies that if stimulus $S$ evokes response $R$ ($S \rightarrow R$), $R$ does not necessarily evokes $S$ ($R \nrightarrow S$); if it does, than a second connection between $S$ and $R$ exists ($S \rightleftarrows R$). Bi-directional (following the associative symmetry hypothesis) implies that if stimulus $S$ evokes response $R$ ($S \rightarrow R$), $R$ automatically invokes $S$ ($R \rightarrow S$); the two associations are inseparable ($S \leftrightarrow R$). According to Anderson and Bower, "[Human associative memory] can perform pair recognition on the bases of either the A to B path or the B to A path" (Anderson and Bower, 1980, p. 220). As a result, current research accepts bi-directionality as a property of human associative memory. The implication here is that for a fact-retrieving task, the roles of stimuli and responses are symmetrical and therefore interchangeable.

Performance of associative memory is inherently hard to measure because of the diversity and incomparability of retrieval tasks (for a discussion, see Anderson and Bower, 1980, p. 153). It is possible, however, to give certain guidelines on how to improve performance in associative memory. An important aspect is the use of preexisting associations. From the definition of paired associative learning (see above), it is obvious that the relationship between the stimulus (or cue) and the response is important for how well people can remember associations (*Ibid.*, p. 189). There is substantial evidence in literature that supports the advantage of using preexisting propositions. As previously mentioned, most of the evidence originates from the field of linguistics. Postman showed that "pre-experimental associative probability [can have] significant effects on learning and retention" (Postman, 1962, p.18) and confirmed this finding in subsequent studies (Postman, Fraser, and Burns, 1968, p. 222).

One of the most famous examples of the use of preexisting associations are mnemonics (see below)

### *Meaning*

As already mentioned, certain associations can carry different meanings and a different amount of meaning for every person. This statement naturally begs the question of how to define and measure meaning. What makes associations meaningful remains a "persistent and controversial problem" (Paivio, 1971, p. 39). Nonetheless, there are several attempts to define and measure meaning.

#### Definition of "Meaning"

Traditional theories define meaning as "some kind of implicit reaction that words arouse, including imagery, nonverbal conditioned reactions, and verbal associative responses" or as "relations between verbal stimuli and overt responses" (Paivio, 1971, p. 40). In these definitions, meaning is just a response to a certain stimulus. Current theories, however, acknowledge the high complexity of the concept of meaning. They add a temporal axis to meaning since "a stimulus [that has] set up a representation […] undergoes a continuous process of transformation as it interacts with the organism and its long-term memory" (*Ibid.*, p. 52). They also acknowledge the existence of multiple levels of "meaning"; whether these levels are continuous or discrete is still debated (see Pylyshyn and Agnew, 1963, versus Paivio, 1971). Paivio distinguishes between three levels of meaning: "the representational process (or representational meaning), referential associative reactions (or referential meaning), and associative chains or structure (or associative meaning)" (*Ibid.*, p. 53). Each of these levels require additional steps of transformation in the owner's mind. Representational meaning "corresponds […] to familiarity in that the familiar has meaning for the individual in the most elementary sense of 'knowing' the stimulus" (*Ibid.*, p. 53). Referential associative reactions evoke stronger responses from people in the sense that they change their behavior or reactions after being exposed to a stimulus (*Ibid.*, p. 57). These two lower levels of meaning incorporate the traditional definition of the term. Finally, associative meaning implies the "development of associative connections or an associative structure involving different referents or conceptual categories" (*Ibid.*, p. 57). These structures can result in an organized system of imagery with spatial organization. They can either be sequential or hierarchical, depending on the nature of the object's referents. This was the first

time that meaning and associations were not only seen as a linear sequence of stimuli and responses but potentially as a non-linear graph. This is important to this research as it shows that there are not many limitations on what kind of associations people can use to evoke a response from a stimulus; as long as people can somehow create an association that is meaningful to them, they will remember it.

Another angle for investigating meaning comes from semiotics, a field in which researchers look at how meaning is transferred between signs and observers. A sign is a combination of a physical expression (the signifier) and an underlying concept (the referent), to which the sign refers. The exact definition of the term "meaning" in semiotics is heavily debated, resulting in some researchers completely rejecting the term (Nöth, 1995). Commonly accepted, however, are the ideas of denotations and designations. A denotation is the initial signified that a sign intends to capture. The signified does not refer to a specific instance but instead to a prototypical category. A designation refers to connotata of a signifier, which are additional referents to the original denotatum. A major difference between denotation and designation is that the latter implicitly includes the interpretation of an observer (Nöth, 1995). The existence of designations allows people to attach connotata, i.e. additional semantic meaning, to real-world objects and other signifiers in general.

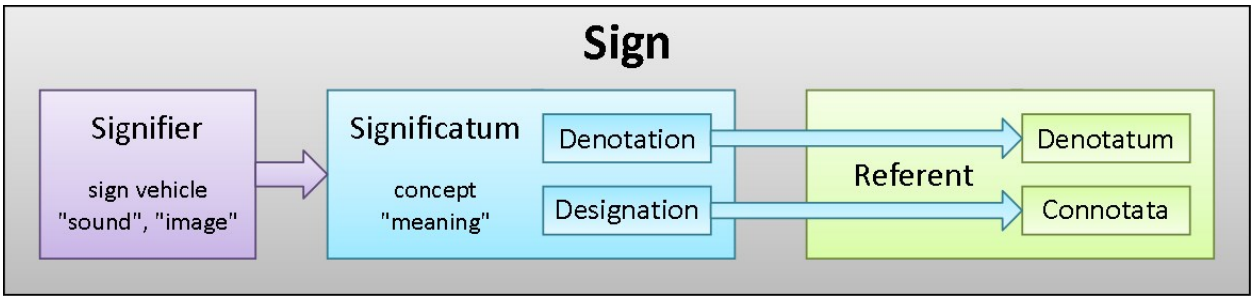



**Figure 20: Peirce's triadic model of signs (adapted from Morris, 1971)**

## Measurement of Meaning

Noble suggested a method for measuring the "amount" of meaning—the meaningfulness—by counting "the average number of continuous written associations given to the item in a standard time period […] by a group of subjects" (Paivio, 1971, p. 45). For Skinner, meaning is less about

the relation between stimulus and response and more about the "conditions under which behavior [stimulus and response] occurs" (*Ibid.*, p. 48). Noble's notion of meaningfulness is very traditional in the sense that he interprets it as a function between an element and a group rather than between an element and an individual, like Skinner does. Although both approaches are valid, I focus on an individual-centered interpretation of meaning.

For this research, the levels (or amount) of meaning and the degree of abstractness (or concreteness) of meaning are two important concepts; the first one refers to the size of the structure of the associative meaning, whereas the second one refers to the relationship between stimulus and response. The level of meaning primarily depends on the complexity of the stimulus (Paivio, 1971, p. 58). Research has shown that an increased level of abstractness of a stimulus makes it less likely to retrieve a response (*Ibid.*, p. 60). Research has also shown that many people give a stimulus the same meaning (*Ibid.*, p. 51). In summary, the relationship between stimulus and response can be called concrete when they share some sort of verbal or imagery connection. While concrete associations generally contain more meaning, some basic abstract associations can do that as well.

In general, there is strong evidence that words and imagery are coded differently in human memory, an effect first described in the dual-coding hypothesis (Paivio, 1971, p. 233).

### *Mnemonics*

One of the earliest associative learning devices are mnemonics—originally called mnemotechnics (Yates, 1966, p. 23). Shortly after Aristotle published his initial associationist theory (Aristotle, 1973, book II, chapter 451b), the first written evidence of the use of associations to improve remembering emerged. In Cicero's book De Oratore (Cicero, 1988, book II, chapter 86), he credited Simonides of Ceos, a Greek poet, with the invention of mnemonics (from Greek *μνήμων* (*mnēmōn*): mindful). Simonides suggested associating certain words or stanzas with locations or areas within a building; walking through the building in one's mind could then enable people to remember the associated words.

Although visual images are powerful cues for association, mnemonics can be of other nature as well (Baddeley, 1998, p. 133): "indeed, during certain historic periods […] visual imagery mnemonics were discouraged, and mnemonics based on meaningful associations regarded as

more acceptable" (*Ibid.*, p. 133). It is possible to formalize mnemonics in the following way: instead of learning a certain meaningless association between stimulus $S$ and response $R$ ($S \overset{meaningless}{\longleftrightarrow} R$), people learn a meaningful one using a mnemonics device $M$ ($S \overset{meaningful}{\longleftrightarrow} M$), and then link the mnemonic device to the actual response through associative chaining ($M \overset{meaningful}{\longleftrightarrow} R$). Instead of having to remember a meaningless assocation, people now only have to remember two meaningful associations; this process increases the amount of meaning and alters its degree of abstractness through associative chaining.

### *Associationism and Full-arm Pointing Interaction*

When a digital system uses real-world proxies for interaction, people have to recall the proxy-object that is associated with the desired interaction. Since proxy-objects are by definition only a representation of the underlying interaction (see 2.2.5), a meaningful semantic association between interaction (stimulus) and proxy (response) can increase learnability, memorability, and selection performance. Here, retrieval of the proxy-object from memory is an example for associative chaining from the initial stimulus (e.g., "It is too dark."), the first response (e.g., "Have to turn on the light."), and the final response (e.g., "Point at the lamp turns it on.").

# Chapter 3 Conceptual Framework for Analyzing Pointing-based Interaction

In Chapter 1, I pointed out the potential advantages of pointing-based interaction over existing techniques for digital artifact selection in smart environments. In this chapter, I first define pointing-based interaction with established vocabulary from the fields of human-computer interaction and cognitive psychology. Then I bring together existing knowledge to describe the details of pointing-based interaction and present a framework that I can later use to analyze and compare room-based interaction with other types of Human-Environment Interaction.

## 3.1 Definitions of Pointing-based Input, Real-world Proxies, *Room-based Interaction*, and *Room Pointing*

Pointing-based input is one particular type of *selection mechanism* (see 2.2.1). "Pointing-based" means that people use mid-air full-arm pointing gestures for system input. One can therefore see pointing-based input as a subset within Bolt's classification of manipulation-based and sign-based gestures (see 2.2.3) or as the selection mechanism for Cockburn's *Air Pointing* interaction techniques (Cockburn et al., 2011, see 2.2.3). A pointing gesture occurs mid-air when it is distal, i.e., when actors do not touch any object with their fingers or use any object to support their arm (see 2.3.2). I consider a pointing gesture to be full-arm when actors use the entire arm (shoulder, upper arm, elbow, lower arm, wrist, hand, and fingers) in the production of the pointing gesture.

Real-world proxies are a particular type of *selection proxy* (see 2.2.1). With real-world proxies, people use real-world objects as interaction proxies (see 2.2.5). From a linguistic and semiotic perspective, the type of selection proxy determines whether a gesture is deictic (pointing toward a real-world object) or emblematic (performing an intrinsically meaningful gesture in personal space), and the type of pointing gesture has an influence on the people's mental model (see 2.3.1).

As people have to interact through an selection mechanism with a real-world object in order to invoke system functionality, pointing-based input and real-world proxies complement each other in the design of an interaction technique. I call techniques that combine pointing-based input and real-world proxies *Room-based Interaction*. The instance of a room-based interaction technique that I am using throughout my research is called *Room Pointing*.

Table 1 sketches out the design space of this dissertation and puts the concepts of selection mechanism and selection proxy the relation to each other. It also shows how existing interaction research papers (*), interaction paradigms (†), and technologies (‡) cover certain areas in this space. The first column lists the type of selection proxies with the nomenclature that I use throughout my dissertation followed by the equivalent term from Cockburn's Air-pointing design framework (APDF, see 2.2.3) and linguistics (Ling, see 2.3.1).

**Table 1: Design space for different combinations of selection mechanisms and proxies**

| **Selection mechanism →**<br>**↓ Selection proxy ↓** | **Full-arm-pointing-based** | **Not full-arm-pointing-based** |
|---|---|---|
| **Real-world proxies**<br>**(APDF: absolute location)**<br>**(Ling.: deictic)** | ***Room Pointing***<br>*XWand**<br>(Wilson and Shafer, 2003) | *Tangible Bits**<br>(Ishii and Ullmer, 1997)<br>e.g., wall-mounted buttons‡ |
| **Screen-based proxies**<br>**(APDF: device-relative)**<br>**(Ling.: deictic)** | *Gyro Point* and *Remote Point**<br>(MacKenzie and Jusoh, 2001)<br>Nintendo Wii Remote‡<br>Microsoft Kinect‡ | PC WIMP†<br>touch interfaces† |
| **Other proxies**<br>**(APDF: e.g., body-relative)**<br>**(Ling.: emblematic)** | *Ray-casting Air-pointing**<br>(Cockburn et al., 2011)<br>*Virtual Shelves**<br>(Li et al., 2009) | *Body Mnemonics**<br>(Ängeslevä et al., 2003) |

As Table 1 shows, pointing-based input can and has been combined with non-real-world proxies, and real-world proxies have been combined with non-pointing-based input. One can use the Air-pointing design framework to map out the design space for potential proxy types, as the concept of selection proxy is related to the concept of *reference frame* in the context of pointing-based input (see 2.2.3). Using "absolute location" as a reference frame is related to using real-world proxies, and using a "device-relative" reference frame is related to on-screen proxy objects (e.g., icons in WIMP interfaces). Interaction techniques with "body-relative" device frames are oftentimes pointing-based (e.g., *Virtual Shelves* by Li et al., 2009), although some touch-based techniques exist (e.g., *Body Mnemonics* by Ängeslevä et al., 2003). I omitted the "object-relative" reference frame from the table because it is of less interest in the context of pointing-based interaction since it is predominantly used in combination with touch as selection

mechanism; see (Cockburn et al., 2011) for an overview and *Marking Menus* (Kurtenbach and Buxton, 1994) for an example.

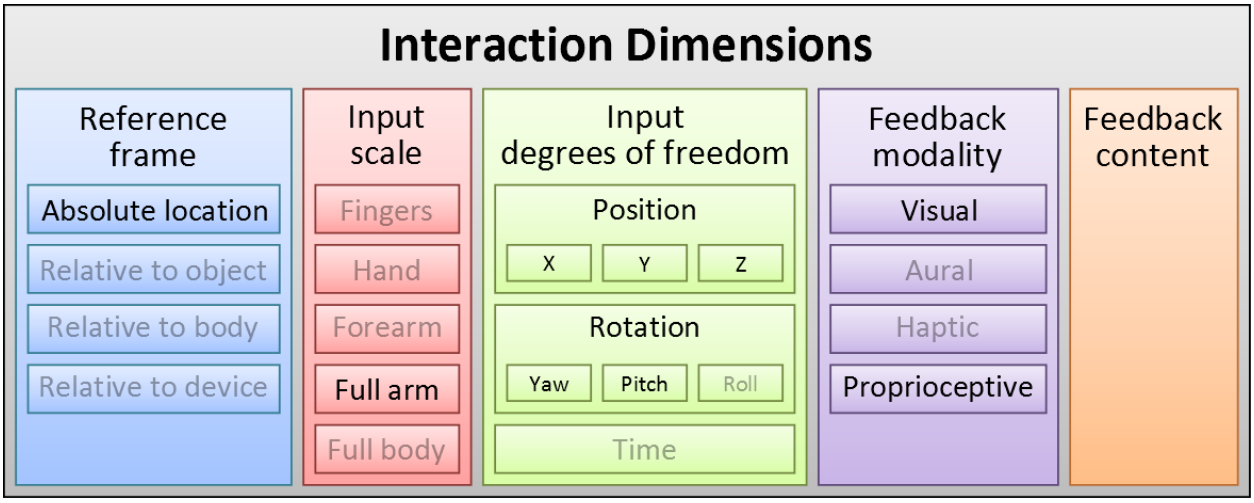



**Figure 21: Room-based interaction in the extended *Air-pointing Design Framework* (adapted from Cockburn et al., 2011, p. 405)**

In summary, room-based interaction is a group of selection techniques that use pointing-based input and real-world proxies for selection. The term "real-world proxies" means that digital artifacts are mapped to real-world objects using the full capabilities of associative memory (see 2.5.5), and interacting with these objects selects the associated digital artifact. The (smart) environment acts as a single-level storage space for selection proxies (see 2.2.5). People can interact with selection proxies by performing a mid-air full-arm pointing gesture toward them (see 2.3.2).

## 3.2 A Framework for Analyzing Pointing-based Interaction Instruments

In this section, I create a framework for analyzing mid-air full-arm pointing gestures and their effectiveness as selection mechanism. The goal of this framework is to enable a comparison and highlight the similarities and differences between three different types of interaction techniques for Human-Environment Interaction: touch-based interaction, mid-air full-arm pointing gestures with real-world proxy-objects, and mid-air full-arm pointing gestures with body-relative proxy-objects.

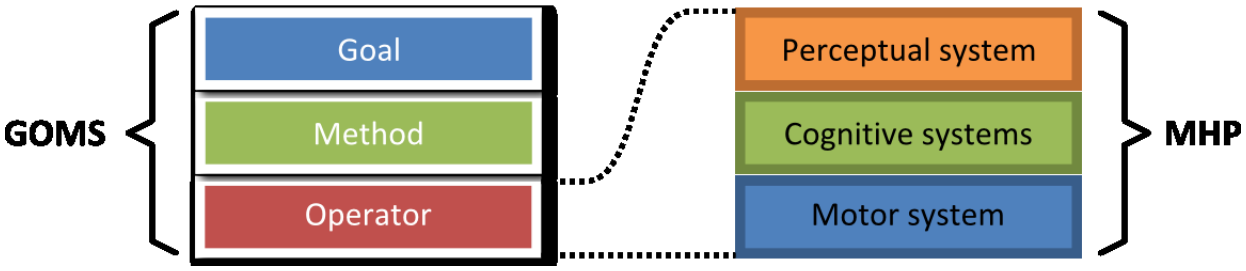



**Figure 22: Components of and legend for the following GOMS / MHP analysis**

Research has shown that people's cognitive understanding of a pointing gesture depends on whether it is toward a real-world object (deictic) or toward a body-relative location (emblematic) (see 2.3.1). Little is known, however, whether these differences in people's mental model influence the performance of an interaction technique. Figure 24, Figure 26, and Figure 29 sketch out the cognitive processes during the creation of touch gesture, a deictic pointing gesture, and an emblematic pointing gesture. The rough structure of the following analyses is based on the GOMS model, the fine structure based on the Model Human Processor (see 2.4.5). See Figure 22 for a reminder of the components of a GOMS / MHP analysis.

This framework is grounded in existing research in (particular sections 2.3 through 2.5) and describes an approximation of peoples' cognitive model when performing abovementioned tasks. The purpose of this framework is informing the hypotheses I use in my users studies. It is not, however, a definitive description of everyone's cognitive model and it is more geared toward novice and intermediate users than expert users. Especially with higher level of expertise, some people might develop very individual cognitive models. While the results of my studies match the predictions from this framework, I make no claim that my results confirm or verify this framework. The reason for this reluctance is that my users studies are designed to investigate room-based interaction and not specifically to verify this framework.

**Figure 23: Feedback-based direct-touch input**

### 3.2.1 An Analysis of a Feedback-based Direct-touch Input Gesture

In this section, I will analyze a typical selection procedure on a

feedback-based direct-touch input device. Performing touch input to a digital system is a task that has many cognitive similarities to reaching and deictic pointing (Marteniuk et al., 1987). I will use "Start Firefox on my smart phone" as an example in my analysis. I assume that the user is already holding the device, and that the device is ready to receive user input, i.e. it is turned on, unlocked, and on its home screen.

***Production of a Direct-touch Input Gesture in the Model Human Processor***

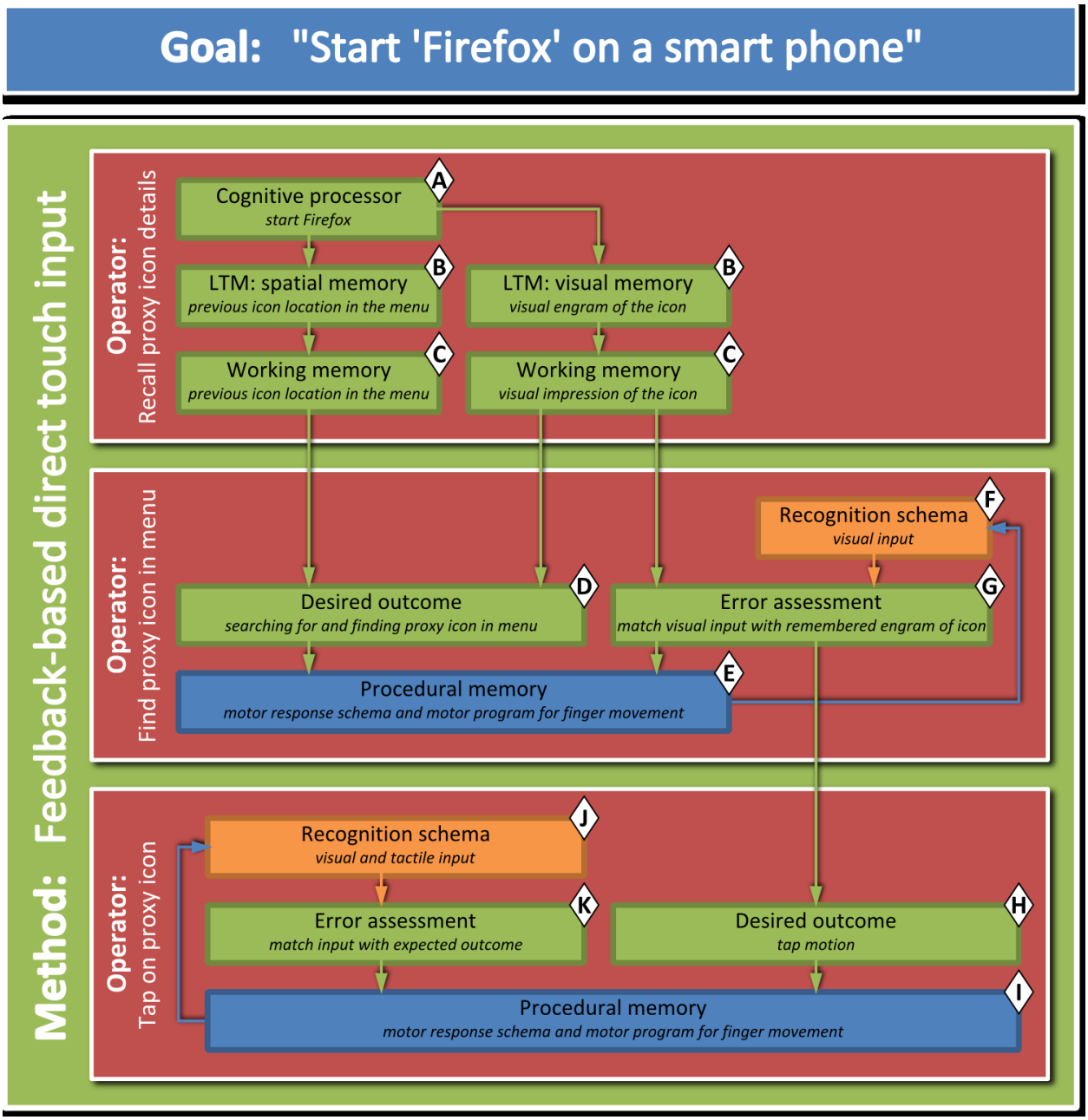



**Figure 24: Cognitive processes during the production of feedback-based direct touch input**

As most actions, the selection starts with a verbal description of the goal: starting Firefox on the smart phone. The first operator of the overall goal is recalling a visual and spatial representation of the goal's verbal descriptor (A – C), the second one is finding the proxy icon within the menu structure (D – G), and the last one is producing a tap gesture on the proxy icon (H – K).

A – C At the beginning of the first operator, people have the verbal descriptor of the overall goal ("start Firefox") loaded in the cognitive processor or central executive (A) (see 2.4.3). Then they retrieve the visual imagery of the icon from visual long-term memory and the last known location of the icon within the menu structure from spatial long-term memory (B). After this, the previously retrieved information is loaded into working memory (C).

D – G In the second operator, people now use both pieces of information to find and visually acquire the icon in the menu structure. First, they formulate the desired outcome of their movements, that is performing a series of swipe gestures that advances the menu to the correct screen (D). Then they calculate the necessary motor program for this movement from an existing and proficiently known motor-response schema for on-surface swiping gestures (E). After this, they evaluate the success of their finding effort (G) by matching current visual sensory feedback (F) to the visual schematic of the Firefox icon now stored in working memory (C) (see 2.4.3). If the error is within acceptable limits, in this particular case: if users have successfully identified the correct icon, they are now ready to produce a tap gesture at the location of the icon.

H – K In the final operator, people use their knowledge of the icon's location to produce the appropriate tapping motion. As with every motoric production, people determine the desired outcome of their tapping gestures based on the location of the icon (H). Then they calculate the necessary motor program for this movement from a well-known motor-response schema for arm, hand, and finger movement (I). After this, they evaluate the correctness of their finger movement (K) by matching visual sensory feedback (J) to the calculated trajectory of their tapping gesture. If the error is within acceptable limits, the stroke phase of the deictic tapping gesture is complete, and people enter the holding or retraction phase of the gesture (see 2.4.4).

### *Analysis of the Production of Direct-touch Input Gestures*

In this section, I analyze the three operators (recall proxy icon, find icon, and perform gesture) required to complete the overall goal.

**First operator:** Recalling the icons visual appearance and former location (B) depends on (semantic) visual memory (see 2.5.5) and spatial memory (see 2.5.3). The visual schematic of the icon is important because it later acts as input in the error assessment process (G); the spatial information is important because it defines a starting position for the—relatively slow—visual search for the icon. Although none of these information are essential, they accelerate the following operator (finding the proxy icon). Without any spatial information, people would have to visually search the entire input space and not just a subsection; without any visual information, people would have to rely on reading and linguistically processing the icon labels or simple guessing the correct icon using relational queues (see 2.5.5). Previous research confirmed that people can recall the visual schematic of icons well, especially when the icons are were designed with people's associative abilities in mind (see 2.5.5). Spatial memory, in contrast, is more expertise-driven than visual and relational memory, i.e. people acquire it more implicitly as a byproduct of interacting with objects (see 2.5.3). That makes spatial memory also more susceptible to failure when object location changes. In HCI, numerous research has shown that spatial stability benefits people's performance with user interfaces (e.g., Gutwin, Cockburn, Scarr, Malacria, and Olson, 2014). Overall, it is reasonable to assume that people can recall the proxy icon reasonably well, as long as it does not change its location.

**Second operator:** Finding the proxy icon within the input space depends on the structure and size of the input space, as well as the people's familiarity with the input space. The structure of the input space can be flat, e.g., the keys on a keyboard or remote control, linear, e.g., the scrollable list menu common in today's smart phones, or hierarchical, e.g., file browsers in desktop operating systems. Complexity analysis describes how the times it takes to find an object depends on the structure of the input space. On a flat input space, this time is equal for all objects: $O(n) = 1$ (e.g., hashtable). On a linear input space, this time depends on the number of elements in the input space as in average half of the elements have to be traversed $O(n) = n$ (e.g., linked list). On a hierarchically structured input space, the structure helps decreasing retrieval time to $O(n) = \log n$ (e.g., tree). These three example are based on data access by a

computing system, and access times might not directly be comparable to that of a human. There are studies, however, that have reported similar performance behavior in humans, e.g., that a flat menu structure has advantages over a hierarchical one (Scarr, Cockburn, Gutwin, and Bunt, 2012). Overall, it is reasonable to assume that navigating the user interface to find the desired proxy icon will take the majority of time for reaching the overall goal and that this time highly depends on people's familiarity with the input space, its structure, and its spatial stability.

**Third operator:** Performing a tap gesture uses a simple and frequently used motor program. Research has shown that people can perform this type of gesture quickly and accurately (Fitts's Law, see 2.2.3).

The conclusion of this analysis is that people should be able to perform the goal of selecting a proxy icon from a menu accurately. The time it takes to make such a selection mostly depends on people's performance in the second operator: finding the correct icon. For this operator, the structure of the input space and people's familiarity with the input space are crucial. This also means that a fundamental improvement in selection time (e.g., from $O(n) = n$ to $O(n) = \log n$) can only be achieved by changing the structure of the input space.

### 3.2.2 An Analysis of a Mid-air Full-arm Pointing Gesture toward a Real-world Proxy Object

In this section, I will give a detailed walk-through of the production of a mid-air full-arm pointing gesture. As previously mentioned, deictic pointing toward a real-world object has many cognitive and motoric similarities to feedback-based direct-touch input (Marteniuk et al., 1987). Therefore, I expect both interactions to be of similar structure and differences in performance to be the result of smaller difference on the operator level. I will use "Watch CNN" as an example in my analysis. In order to change the station to CNN, the user has to make a pointing gesture toward the door. A mnemonic device to remember this somehow abstract association could be "CNN is my door to the world".

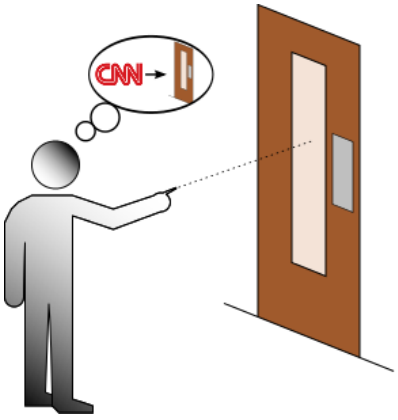



**Figure 25: Room-based interaction with deictic pointing gestures**

***Production of a Deictic Pointing Gesture in the Model Human Processor***

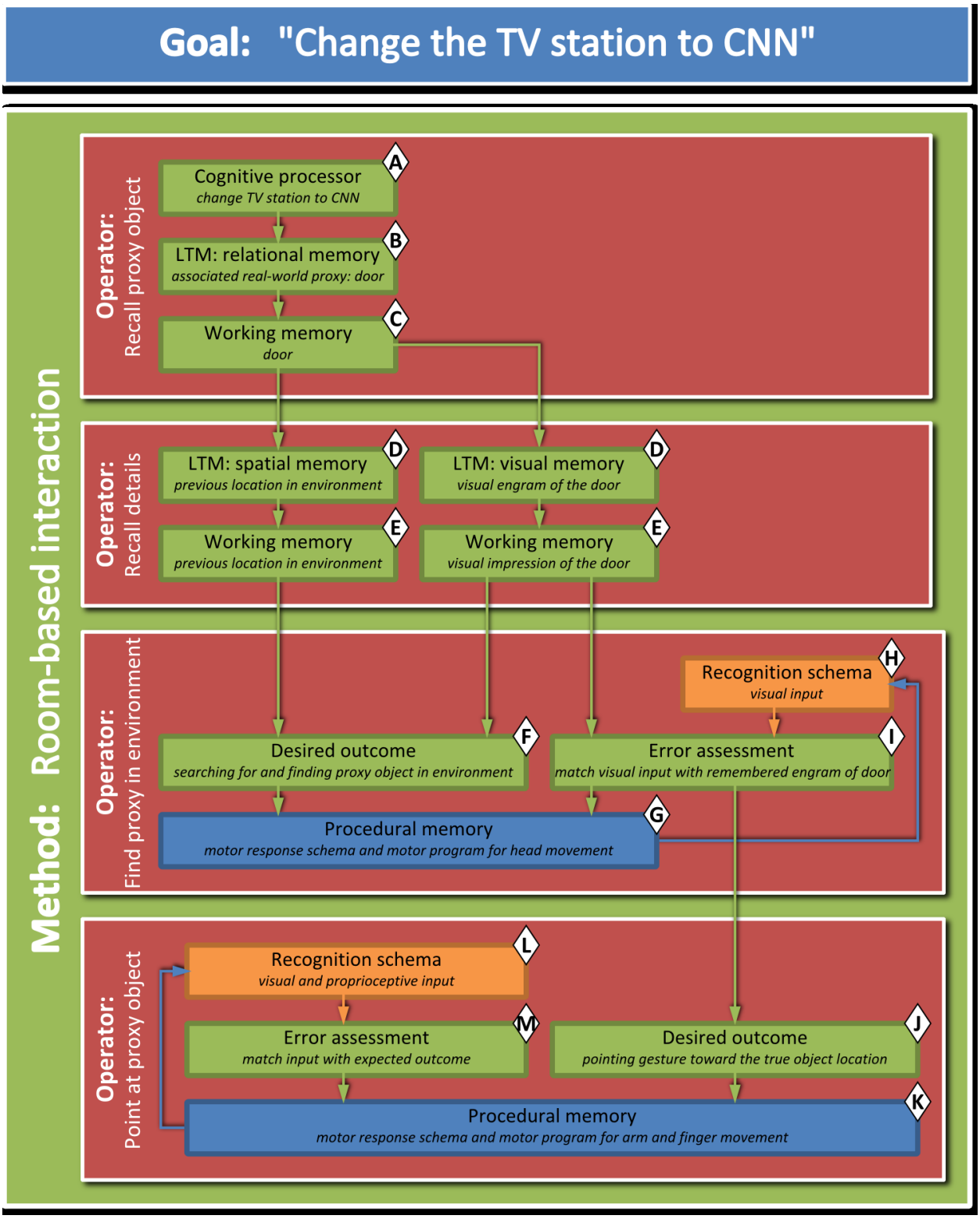



**Figure 26: Cognitive processes during the creation of a deictic pointing gesture**

As most actions, the selection starts with a verbal description of the goal: changing the TV station to CNN. The first operator of the overall goal is recalling the real-world proxy object that is associated with the goal's verbal descriptor (A – C), the second one is recalling the spatial and visual details about the proxy object (D – E), the third one is finding the proxy object in the environment (F – H), and the last one is producing a deictic pointing gestures toward the real-world proxy object (J – M).

A – C At the beginning of the first operator, people have the verbal descriptor of the overall goal ("change TV to CNN") loaded in the cognitive processor or central executive (A) (see 2.4.3). They are then retrieving the real-world proxy object that is associated with the channel "CNN" from relational memory (B). After this, a verbal descriptor of the proxy object ("door") is loaded into working memory (C).

D – E In the second operator, people now try to retrieve the visual imagery of the real-world object from visual long-term memory and its last known location in the environment from spatial long-term memory (D). These two information are then loaded into working memory (E).

F – I In the third operator, people now use both pieces of information for finding and visually acquiring the real-object in the environment. This is the first operator in which motor skills are required. First, people formulate the desired outcome of this of their head-movement: finding and fixating the real-world proxy object (F). Then they calculate the necessary movement pattern and motor programs from a well-known motor-response schema for head movement (G). After the body-movement, people evaluate its correctness (I) by matching visual sensory feedback (H) to the stored visual engram of the real-world object (E). If the error is within acceptable limits, people are now have successfully fixated the real-world object and are ready to produce the deictic pointing gesture toward the door.

J – M In the final operator, people use the knowledge of the object's location to produce the appropriate pointing gesture toward the real-world proxy object. First, they determine the desired outcome of their pointing gestures based on the location of the door (J). Then people calculate the necessary motor program for this movement from a well-

known motor-response schema for upper body, shoulder, arm, hand, and finger movement (K). After this, they evaluate the correctness of their movement (M) given by matching visual sensory feedback (L) to the perceived direction of their pointing gesture. If the error is within acceptable limits, the deictic pointing gesture is complete, and people enter the holding phase of the gesture.

### *Analysis of the Production of a Deictic Pointing Gestures*

In this section, I analyze the four operators (recalling real-world proxy object, recalling its spatial and visual details, finding the object in the environment, and performing a deictic pointing gestures toward the object) required to complete the overall goal.

**First operator:** Association of a digital artifact with a real-world proxy object is one of the distinctive traits of room-based interaction, and remembering this association and recalling it in a fast and accurate manner are crucial parts of using room-based interaction successfully. Recalling the real-world proxy object depends entirely on relational memory. Research has shown that people's relational memory can function exceptionally well if there is a meaningful connection ("CNN is my door to the world") between the stimulus ("Watch CNN") and response ("door") (see 2.5.5). Existing semantic memory can be of great help for remembering new associations between two objects or concepts, and people can also easily fabricate associations between seemingly unrelated items (see 2.5.5). Overall, evidence from existing research lead me to the conclusion that people should have little problems remembering associations between system commands and real-world proxy objects.

**Second operator:** People's performance in retrieval of the visual imagery of the real-world object (C) depends on their familiarity with the object, and retrieval of the last known spatial location (C) depends on their familiarity with the environment. Given the context of my work, that is interacting in domestic environments, I assume that people know the environment and the real-world objects within well and thus can retrieve the spatial and visual information of an object (i.e., how an object looks and where it is) precisely enough for finding the object in the third operator (F – I ).

**Third operator:** The time and effort people need for locating and fixating a real-world object in the environment (F – I) depends on several factors. Two important ones are the spatial stability

of the environment, i.e. how frequently objects change their location, and the visual distinguishability of an object, i.e. how many similar-looking objects are in the environment. With domestic environments, I assume that proxy object are unique and static. Unique means that people would not confuse an object with another similar-looking one, and static means that the object is not moved around. This can mean that people might have to be careful when deciding which real-world object to use as proxy for a digital artifact. Overall, people should be able to find and visually fixate proxy objects quickly, reliably, and without much effort.

**Fourth operator:** When performing a full-arm pointing gesture toward the real-world proxy object, people's performance depends on the size of the pointing target, i.e. the angular size of the real-world proxy object, and the available feedback channels. As long as the angular size of the pointing target remains above human pointing errors (see 2.4.4), people should not have difficulties to accurately point at a real-world object. Research has shown that people can point most accurately when they can employ visual feedback during the error assessment (M) (see 2.4.4). Overall, people should be able to perform quick and sufficiently accurate pointing gestures toward real-world objects.

The conclusion of this analysis is that people should be able to use room-based interaction. One finding is that pointing at a real-world proxy object is cognitively rather similar to using a feedback-based direct-touch technique. The major difference is that room-based interaction requires an additional operator (the first operator): recalling the association between the system command and the real-world proxy object. Another more subtle difference is that the motor movement during the final operator for touch input (tap motion) is probably faster and more accurate than for room-based interaction (full-arm pointing gesture). In contrast to touch input, room-based interaction offers, however, the possibility for eyes-free interaction. Touch input relies on visual input during the second and third operator (finding and taping on proxy icon) (see 3.2.1). In room-based interaction, people can solely rely on spatial input for finding the proxy object in the environment (third operator) because of the detailed spatial understanding that people have of the environment (see 2.5.3) and solely rely on proprioceptive feedback when pointing at the real-world proxy object (see 2.4.4). As a result, people can skip the third operator (visually fixating the proxy object) altogether. The following full-arm pointing gesture during the final operator is then, however, only guided by proprioception, which should lead to a decrease

in pointing accuracy (see 2.4.4). Figure 27 illustrates the difference in the cognitive processes when using room-based interaction eyes-free, i.e. without looking at the real-world proxy object.

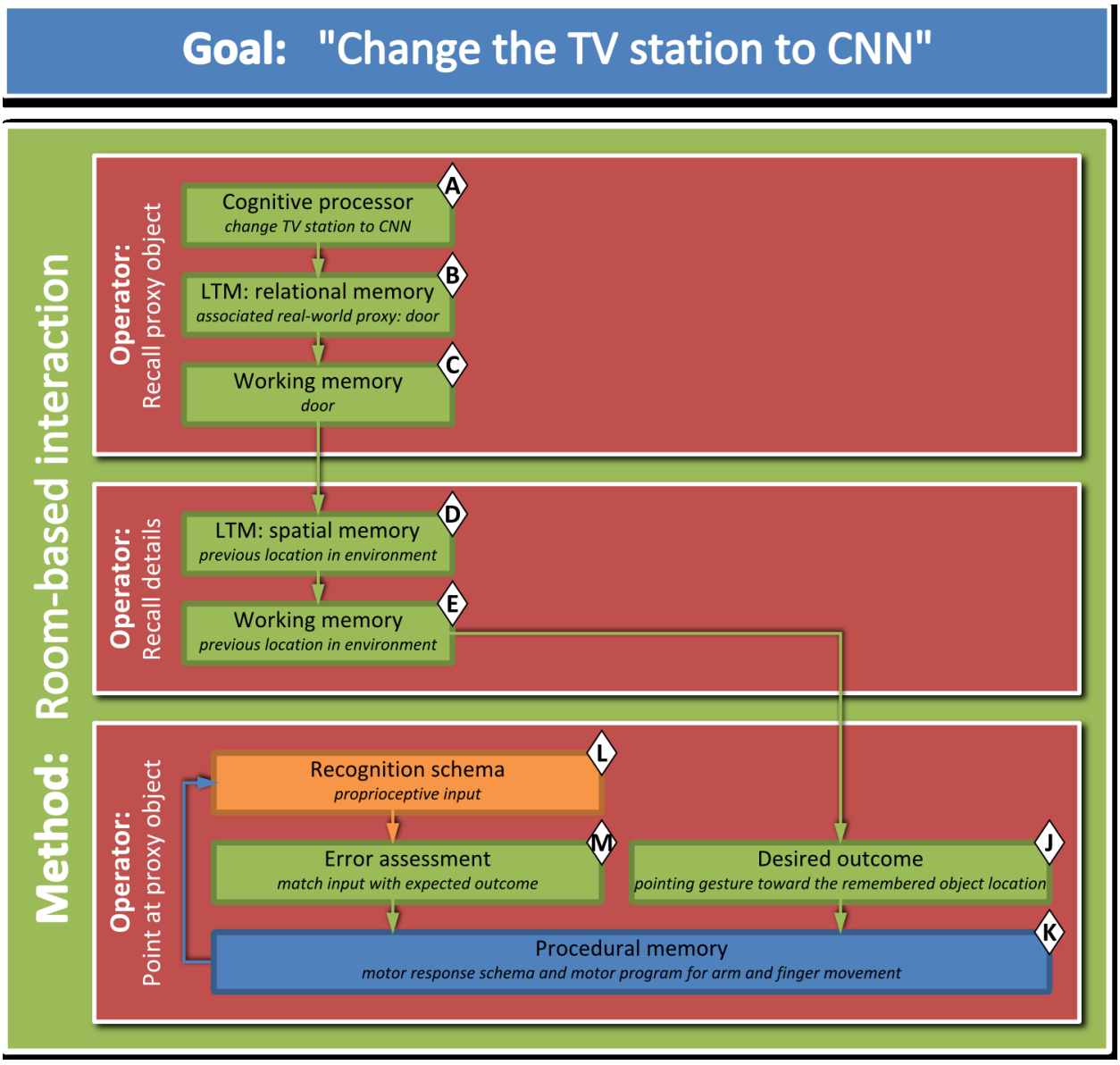



**Figure 27: Cognitive processes during the eyes-free creation of a deictic pointing gesture**

This analysis also gives an explanation about the desired characteristics of real-world proxy objects. Ideally, proxy objects should have some meaning to the user as this allows creating a meaningful association between digital artifact and proxy object, which in turn helps remembering the association (first operator) (see 2.5.5). The meaning between digital artifact

(stimulus) and proxy object (response) should also be unique, so that there are only a few possible responses to a given stimulus (first operator) (see 2.5.5). The location of proxy objects should be static, so that the users do not have to engage in a time-consuming visual search of the entire environment (second operator). Last, the proxy object should be visually distinct from other real-world objects, so that users do not confuse proxy objects.

### 3.2.3 An Analysis of a Mid-air Full-arm Pointing Gesture toward a Body-relative Proxy Zone

In this section, I will analyze a typical selection procedure using emblematic mid-air full-arm pointing gestures. While the difference between deictic gestures toward real-world proxy objects and emblematic gestures toward a body-centric region in space might initially not be obvious, a GOMS / MHP analysis is likely to reveal some crucial differences because memory and feedback systems are used differently (see 2.4.2). The example I am using for this analysis is the system command “mute sound system”. In order to issue this command to the smart environment, the user has to make a pointing gesture “raise arm slightly, 30° to the right”.

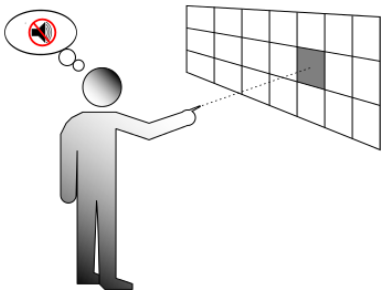


**Figure 28: Body-centric interaction with emblematic gestures**

***Production of an Emblematic Pointing Gesture in the Model Human Processor***

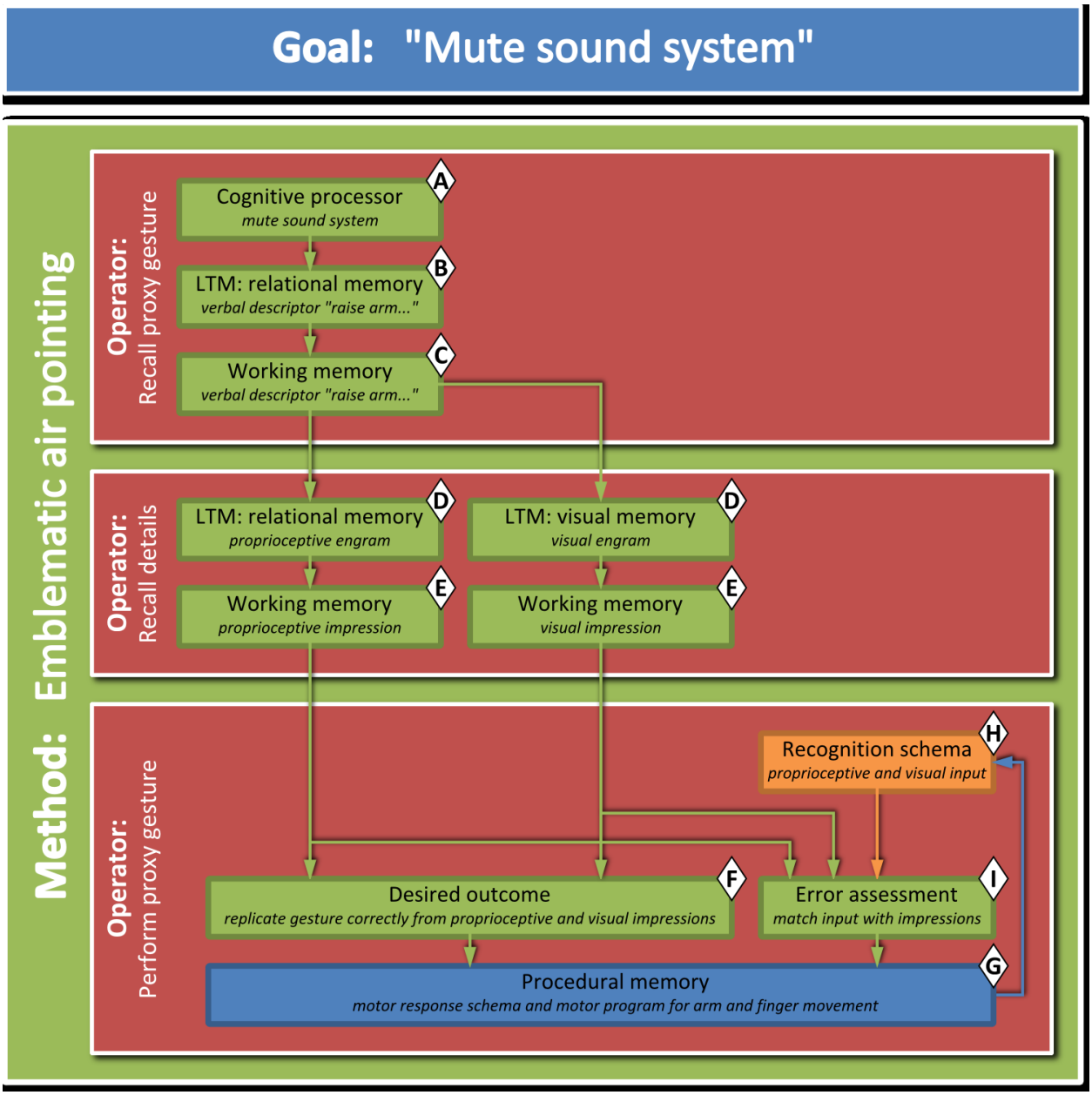



**Figure 29: Cognitive processes during the creation of an emblematic pointing gesture during the cognitive phase of motor skill learning**

As most actions, the selection starts with a verbal description of the goal: muting the sound system. The first operator of the overall goal is recalling a verbal descriptor of the pointing gesture (A – C), the second one is recalling a visual and proprioceptive representation of the pointing gesture and the verbal descriptor (D – E), and the third operator is creating and performing the emblematic pointing gesture (F – I).

A – C At the beginning of the first operator, people have the verbal descriptor the overall goal ("mute sound") loaded in the cognitive processor or central executive (A) (see 2.4.3). Then they retrieve the verbal descriptor of the movement ("raise arm slightly and point approximately 30° to the right") from long-term memory (B). After this, they load the previously retrieved descriptor into the working memory (C).

D – E In the second operator, people then use the verbal descriptor to derive estimates for proprioceptive and visual information about the arm motion (D). Finally, this information is loaded into working memory (E).

F – I In the third operator, people now use the visual and proprioceptive impressions for performing the emblematic pointing gesture. First, they determine the desired outcome of their pointing gestures based on the expected proprioceptive and visual sensory response (F). After this, they calculate the required motor response for body, arm, hand, and finger movement that will bring their body into a position where the expected proprioceptive and visual sensory response can be achieved (G). Finally, people match remembered (E) and actual (H) sensory input and assess whether they are similar enough to complete the gesture (I).

### *Analysis of Emblematic Pointing Gestures*

In this section, I analyze the three operators (recalling proxy gesture, recalling its proprioceptive and visual details, and performing a emblematic proxy gestures) required to complete the overall goal.

**First operator:** Relational memory plays a crucial part in recalling the proxy gesture as its role is translating between stimulus ("Mute sound") and response ("raise arm slightly and point approximately 30° to the right"). With emblematic pointing gestures, the response is a verbalization of a motor procedure (see 2.3.1 and 2.5.4). Research has shown that procedural and relational memory are fundamentally different and somehow incompatible, which makes transferring information between them difficult (see 2.5.4). As a result, the verbal descriptor will inherently lack precision. Given the limited associative descriptiveness of the verbal descriptor, the amount of meaning between digital artifact and the verbal descriptor might be relatively low. This in turn would make remembering the connection between system command and verbal

descriptor difficult (see 2.5.5). Overall, it is reasonable to assume that people might have difficulties remembering the association between digital artifact and verbal descriptor and that the remembered verbal descriptor might lack precision.

**Second operator:** People now have to retrieve proprioceptive and visual impressions based on the previously recalled verbal descriptor. Generally, the imprecise nature of the descriptor makes this retrieval difficult, especially for the proprioceptive impression. As mentioned above, proprioceptive information is inherently incompatible with associative memory (see 2.5.4), and people have difficulties storing and retrieving this kind of information from relational memory (see 2.5.4). As a result, the precision of the proprioceptive impression, which is important for creating and assessing the produced pointing gesture, is further reduced. Overall, it is again reasonable to assume that people might have difficulties retrieving precise visual and proprioceptive information.

**Third operator:** People perform the emblematic pointing gesture toward a body-relative proxy zone. Performance in this step mostly depends on the precision of the recalled proprioceptive and visual impressions, as they are used for both calculating limb movements as well as assessing the accuracy of the produced pointing gesture. Since these impressions are inherently imprecise, I assume that people's accuracy will be generally low.

My conclusion of this analysis is that I expect people to have problems producing accurate emblematic pointing gestures, and that interaction techniques that use these kind of gestures will show low performance. This conclusion is backed up by existing research that has pointed out people's difficulties in acquiring procedural memory and motor proficiency (see 2.5.3). One interesting aspect of using procedural memory, however, is that people's performance significantly increases when reaching higher levels of proficiency (associative and autonomous phases, see 2.5.4). In the autonomous phase, semantic memory (B – E) is bypassed, and the cognitive processor triggers procedural memory directly. In the context of my research, the autonomous phase is of little interest, however, since it takes more practice to reach this phase than a one-hour experiment offers.

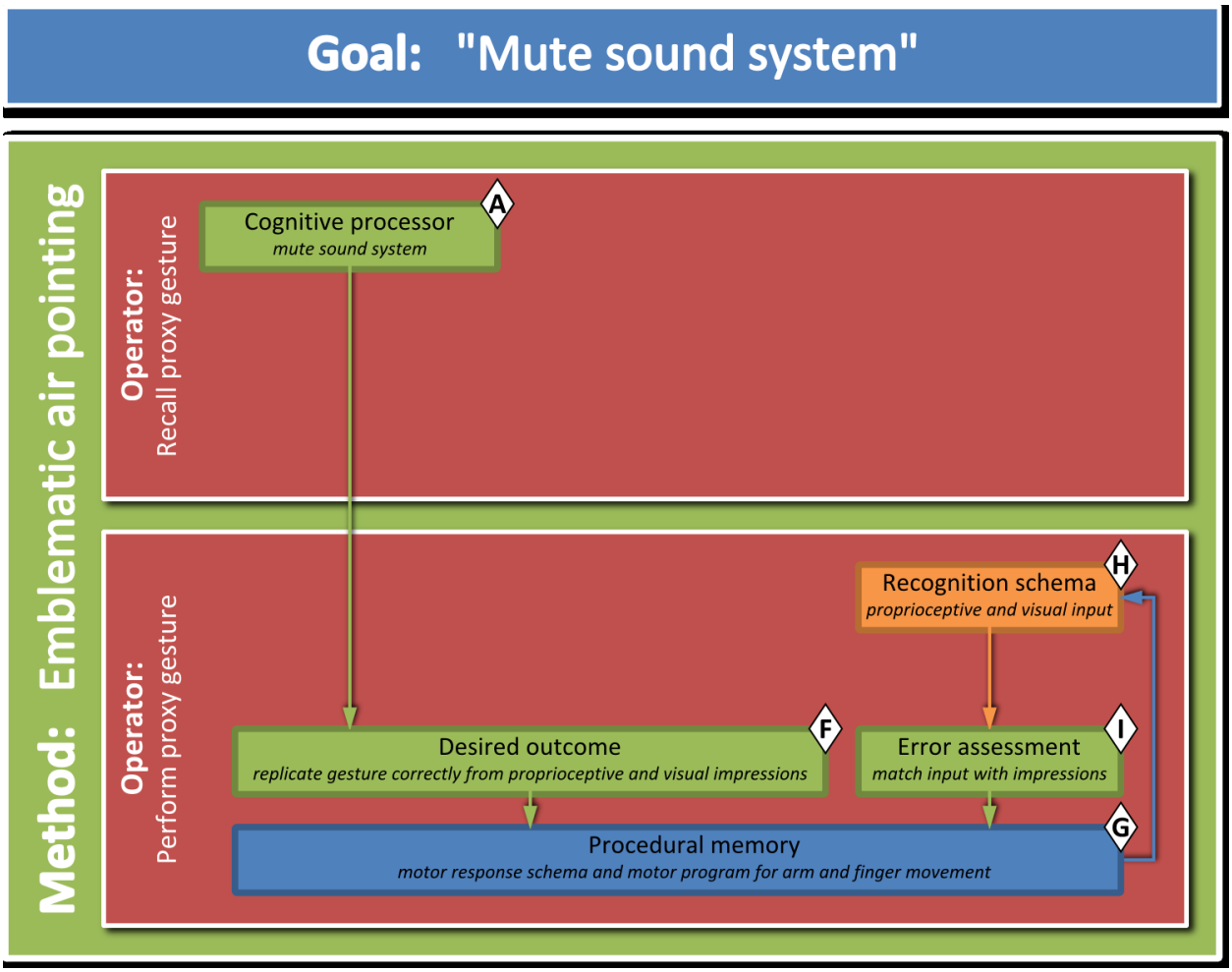



**Figure 30: Cognitive processes during the creation of an emblematic pointing gesture during the autonomous phase of motor skill learning**

### 3.2.4 Summary and Conclusion

The comparison of the cognitive processes in the three different HEI-techniques—feedback-based direct-touch, mid-air full-arm pointing gestures toward real-world proxy objects, and mid-air full-arm pointing gestures toward body-relative proxy zones—revealed several similarities as well as some crucial differences.

#### *Feedback Channels*

The combination of selection mechanism and selection proxy in an interaction technique usually determines its main feedback channel. In general, feedback plays an important role in improving the accuracy of an interaction technique. Touch interfaces and *Room Pointing* are similar in that

they both use visible objects as selection proxies: on-screen icons and real-world objects. This means that both techniques rely on vision as the main feedback channel. For both techniques, vision is necessary in order to find the proxy icon and accurately tap on or point toward it. Despite this similarity, I expect that people will show higher accuracy in touch interface due to a difference in selection mechanism: touch interfaces start with comparably accurate proximal pointing and transition to touch, which adds haptic feedback, whereas *Room Pointing* only uses comparably inaccurate distal pointing. *Room Pointing*, however, offers people the possibility to blindly point toward the real-world proxy object, thus rely purely on proprioception. This is possible because people have an accurate spatial model of objects in familiar environments and a good understanding about proprioceptive feedback when performing deictic pointing gestures. In *Virtual Shelves*, in contrast, selection proxies are virtual and thus do not generate visual feedback directly. Instead, people have to rely on feedback generated from the selection mechanism—the mid-air full-arm pointing gesture—, which is mostly proprioceptive and, to a lesser degree, visual. As discussed above, I expect low selection accuracy in *Virtual Shelves* due to the initial inaccuracy of proprioceptive feedback. Despite using the same selection mechanism, I expect people's selection accuracy in *Virtual Shelves* to be lower than in *Room Pointing* because the different selection proxies in both techniques result in different main feedback channels.

### *Memory Systems*

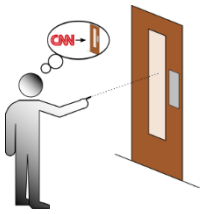
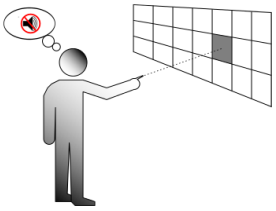
The combination of selection mechanism and selection proxy also determines which memory systems are most relevant for an interaction technique. In general, the involved memory systems and the pre-existing knowledge plays an important role in the initial learnability of an interaction technique. As with feedback, touch interfaces and *Room Pointing* are similar in that they both mostly rely on semantic memory, i.e., spatial, relational, and visual memory. With touch interfaces, people require a single cue–response pair for translating their intention to the visual and spatial information about the proxy icon. With *Room Pointing*, people need two pairs: one for translating the intention to the proxy object and another one for retrieving the object's visual and spatial information from memory. The first step in *Room Pointing* is therefore additional compared to touch interfaces. Weather this additional indirection in *Room Pointing* will lead to decreased selection accuracy and increased selection time will most likely depend on the amount of meaning between the intention and the real-world proxy-object. In *Virtual Shelves*, the situation is more complicated as the involved memory systems vary depending on a person's

learning stage (cognitive, associative, and autonomous). During the cognitive stage, both semantic and procedural memory are involved in creating a pointing gesture toward a target zone. This translation between memory systems adds an element of inaccuracy to the execution of the pointing gesture, which I believe will lead to a decreased pointing accuracy.

### *Limiting Factors, Predicted Performance, and Conclusion*

I expect that people will show high selection accuracy with direct touch but might display low selection speed. Whether selection speed will be low will most likely depend on the structure of the input space: on a flat input space, which does not require menu navigation, selection speed will be high, on a hierarchical input space, it will be low. I expect people to show high selection speed with *Room Pointing*. Selection accuracy, however, might be reduced due less accurate feedback compared to direct touch. For *Virtual Shelves*, I expect people to show the same level of selection speed as in *Room Pointing* due to the similarities in selection mechanism. However, I expect significantly lower (initial) selection accuracy because of the differences in feedback.

**Table 2: Comparison of Direct Touch, *Room Pointing*, and *Virtual Shelves***

| | **Direct touch on on-screen icons: "Touch interface" (3.2.1)**  | **Mid-air full-arm pointing gestures toward real-world proxy-objects "Room Pointing" (3.2.2)**  | **Mid-air full-arm pointing gestures toward body-relative proxy-zones "Virtual Shelves" (3.2.3)**  |
|---|---|---|---|
| **Selection mechanism** | Direct touch | Mid-air full-arm pointing gestures | Mid-air full-arm pointing gestures |
| **Selection proxy** | On-screen objects (icons) | Real-world objects | Virtual zones |
| **Main feedback** | Visual | Visual | Proprioceptive |
| **Secondary feedback** | Haptic<br>Proprioceptive | Proprioceptive | Visual |

| | | | |
|---|---|---|---|
| **Main memory system** | Semantic | Semantic | Procedural |
| **Reference frame** | Device-relative | Absolute | Body-relative |
| **Main limiting factors (selection accuracy)** | — | Pointing accuracy | Gesture recall<br>Pointing accuracy |
| **Main limiting factors (selection speed)** | Menu navigation | — | — |
| **Predicted accuracy** | High | Medium | Low |
| **Predicted speed** | Medium | High | High |

## 3.3 Smart Environments

### 3.3.1 A (Re-) Definition of *Smart Environment*

Since Weiser's initial vision of UbiComp (see 2.2.2), researchers coined several other terms that all refer to different aspects of a smart environment, such as Pervasive Computing, Ambient Intelligence, Smart Environments, and Internet of Things (see 2.1). For my dissertation, I felt that none of the existing terms captured the context of my work properly and that using these terms could mislead readers. This is why I decided to provide my own definition of *smart environment*: a confined physical space with digital artifacts in it. The only assumptions I make about a smart environment are that its dimension (size and height) is typical for a domestic room (e.g., kitchen and living room) or an office space (e.g., offices or cubicles) and that people can control digital artifacts remotely, i.e., through a common digital system.

### 3.3.2 Example Tasks for Command Selection in Smart Environments

An important premise when comparing different interaction techniques for smart environments is that there will not be a single technique that is superior in all potential scenarios. For this, smart environments provide too many different use cases. I want to outline three mundane scenarios in domestic environments to demonstrate this diversity:

1. selecting a movie to watch on TV
2. checking a cooking recipe
3. turning on the living room lights while reading a book

These three scenarios occur daily in homes around the world and feature a non-computer-based main goal or primary task and a UbiComp-based supporting task. When comparing scenarios, I focus on the following aspects: how does the UbiComp interaction fit into the users' process of reaching their main goal, how complex is the UbiComp interaction, and how large is the input, output, and feedback space? These scenarios also represent the primary tasks that I am focusing on in my dissertation: single action selections. All three tasks can be completed with a single selection.

In the dissertation, I specifically limit HEI to artifact selection, which is choosing a single artifact from a larger group

### *Selecting a Movie*

When watching a movie is the user's main goal, the supporting UbiComp task of selecting the movie is rather complex as it requires to make a selection between potentially thousands of digital artifacts. All these artifacts have to be displayed, and users need means for browsing or searching. This complexity demands large input and output space and could make the supporting task disruptive of the primary task. The supporting task, however, occurs serial to the primary task because it happens before the primary task and does not coincide with it. As a result, the cost of interruption is relatively low, and the interaction does not have to focus on selection speed.

The user interface should be able to support complex interactions to accommodate the large input and output spaces. Possible solutions are hierarchical menus or a search function, which require display space for output and text entry for input. As a result, a smart phone or tablet would be a suitable interaction device.

### *Checking a Recipe*

The supporting UbiComp task of checking a cooking recipe is simple as it might only require basic functionality, such as scrolling and zooming. Input and output space is therefore smaller than in the example above. The task can, however, overlap with the primary task (cooking) and can occur multiple times. This co-occurrence forces users to switch contexts between the primary task and the supporting task. These context switches can be very time-consuming and disruptive because cooking includes handling fatty, sticky, and potentially pathogenic ingredients that can

damage touched device or harm people. Device-free interaction would clearly reduce the cost of interruption.

The user interface only has to support a small input space, while providing some means for outputting text and imagery. Given these requirements, smart phones might not be the best-suited interaction device. Instead, a wall-mounted display with gestural input might be more useable.

### *Turning on Lights*

Turning on the lights while reading a book is a minimalistic task with small input space and no system feedback. Readers might be deeply immersed in a book when the need for more light arises in order to proceed with the primary task (reading). Given the simple and brief nature of the supporting UbiComp task, it should be easily executable without requiring a prolonged context switch as this greatly disrupts the primary task.

The user interface should reflect the simplicity and the potentially high cost of interaction. Having to interact with a screen, either hand-held or wall-mounted, in order to turn on the light would majorly disrupt people in reading their book, the primary task. Screen interaction requires full visual and cognitive attention; people would have to complete two context shifts for an interaction of negligible complexity. As a result, a device-, eyes-, and feedback-free interaction, such as a room-based interaction, would be preferable in this particular scenario.

**Table 3: Comparison of UbiComp interaction scenarios**

| | Input space | Output / feedback space | Cost of interruption |
|---|---|---|---|
| **Movie selection** | Large | Large | None |
| **Cooking recipe** | Medium | Medium | High |
| **Switch on lights** | Small | None | High |

These three scenarios clearly show that the requirements for interactions in smart environments can be so different that no single interaction technique will satisfy all of them: every technique has their individual strengths and weaknesses. While touch-based techniques will most likely keep their place in HEI, there are certain scenarios where people might prefer device-, system-feedback-, and eyes-free techniques. With room-based interaction, I present an interaction

paradigm that covers some of the scenarios where touch-based techniques have some shortcomings. In this sense, room-based interaction supplements existing techniques for HEI rather than replacing them.

## 3.4 The Scope of my Research

My research is focused on investigating pointing-based selection mechanisms and selection proxies based on real-world objects.

In my first study (see Chapter 5), I will compare room-based interaction and remote pointing, two types of interaction techniques that use a pointing-based selection mechanism, with touch interfaces. Remote pointing is a common method for interacting with screen-based digital systems from a distance (e.g., Nintendo Wii Remote and Microsoft Kinect), and touch-based interaction is oftentimes considered the default for interacting with smart environments (Ballagas, Borchers, Rohs, and Sheridan, 2006). The purpose of this study is to compare pointing-based interaction with today's touch-based interaction.

In my second study (see Chapter 6), I will compare two interaction techniques that both use pointing-based selection mechanisms but different selection proxies. *Room-based interaction* uses real-world proxies as selection proxies, whereas *Ray-casting Air-pointing* uses body-relative pointing directions. Research has shown that amongst pointing-based selection mechanisms, both approaches for selection proxies (real-world proxies and body-relative proxies) appear to be promising. Given the theory on human memory system, in particular the capabilities of human associative memory, I expect real-world proxies to have an advantage over body-relative proxies in term of memorability and ease of learning (see 2.5.5).

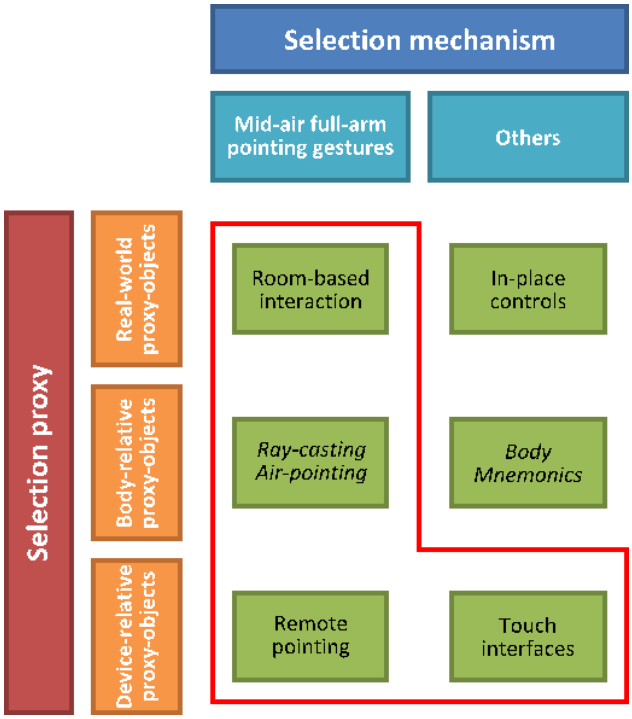



**Figure 31: Design space of my dissertation (red)**

# Chapter 4 The Technical Feasibility of Room-based Interaction

In Chapter 3, I presented the theoretical foundations of room-based interaction and laid out its potential strengths and weaknesses based on previous research in psychology and kinesiology. In this chapter, I describe the implementation of a system that is capable of tracking, processing, and interpreting human pointing gestures. This system requires two separate components: hardware for capturing people's arm, wrist, and hand motions, and software for processing the motion-capture input, calculating pointing direction, and determining selected real-world objects.

## 4.1 Tracking Hardware

There are multiple technologies for tracking the location and orientation of objects and people in three-dimensional space. In my research, I used electromagnetic and optical trackers.

### 4.1.1 Electromagnetic Tracking Hardware

Electromagnetic trackers use an electromagnetic field for determining the location and orientation of tracked objects. As all fundamental vector fields, the electromagnetic field is defined in each location by its energy and direction. Typically, electromagnetic trackers use require three components: a source, a sensor, and a system unit.

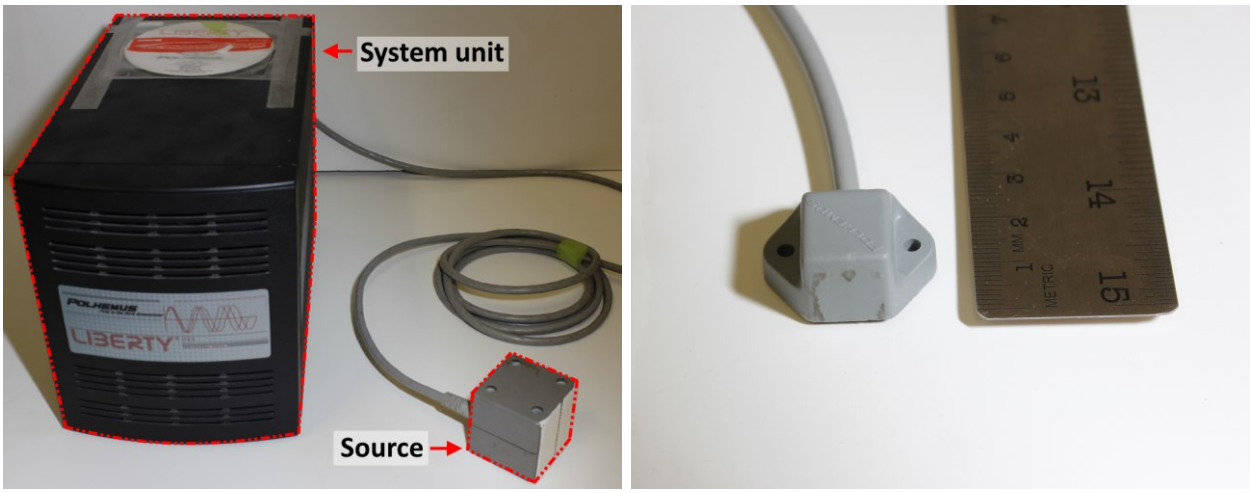



**Figure 32: Polhemus Liberty system unit and source (left) and sensor (right)**

The source is located at the center of the tracked volume and emits an electromagnetic field with a certain strength (see Figure 33 for an visualization). The sensor is attached to each of the

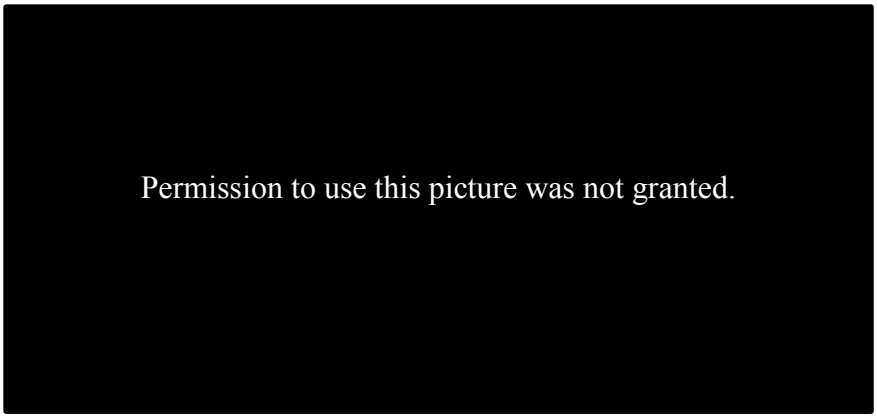



**Figure 33: Electromagnetic field emitted by an electromagnetic tracker (courtesy of Polhemus, Inc.)**

tracked objects and measures the energy and direction of the emitted field. This combination is unique to every location within the tracked volume, thus allows for calculating the sensor's location and orientation.

A general problem of using electromagnetic trackers is that numerous sources, such as electric currents running through wires, metallic objects, and large volumes of water, distort or dampen the electromagnetic field. This means that tracking accuracy is high in close vicinity around the source but then drops off quickly. My initial investigation in the use of the Polhemus Liberty electromagnetic tracker showed that it produced reliable orientation data only within approximately 1 $m$ distance from the source. As a result, electromagnetic trackers might not be best suited for tracking people in domestic environments or similar large tracking volumes.

### 4.1.2 Optical Tracking Hardware

Optical trackers use an array of cameras that are located around a region in space, the tracking volume. The cameras record a video stream and use thresholding to convert it into black and white. Although it is possible to use visible light ($390\ nm < \lambda < 700\ nm$), most current tracking systems use infra-red (IR)

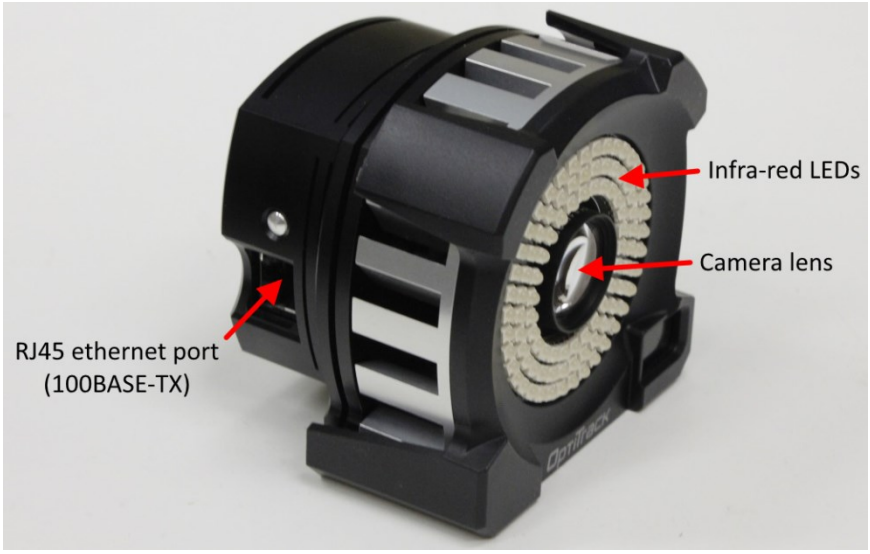



**Figure 34: NaturalPoint OptiTrack S250e camera**

sensitive cameras ($800\ nm < \lambda < 1\ mm$) since IR-light does not interfere with natural light reflected from most objects in the environment or people's cloths. There are, however, two disadvantages to this approach. First, all tracked objects must either be IR-reflective by themselves or IR-reflective tags must be attached to them. Second, the scene must be artificially illuminated since ambient IR-levels are too low. Most dedicated tracking cameras have therefore a set of IR-LEDs to illuminate the environment. These LEDs, however, are so bright that they can cause reflections on smooth and polished surfaces, such as floor tiles; cameras should thus be placed carefully in order to minimize the effect of these reflections. Given the imprecise nature of cameras, it is recommended to use more than the minimum of two cameras (stereoscopic tracking); in general, having multiple cameras mounted so that they capture the tracking volume from different directions leads to more accurate tracking.

For all my studies, I used NaturalPoint OptiTrack S250e IR-cameras (Figure 34). These cameras operate at a wavelength of $800\ nm$ and offer $56°$ field of view; they have up to $250\ Hz$ sampling rate and $4\ ms$ latency. The cameras were connected to my experiment computer through two network switches via 100BASE-TX Ethernet; the switches also powered the cameras through PoE (Power over Ethernet). Depending on the volume I had to cover, I used between six and eight cameras, which were mounted on two-section articulated arms and attached to the lab's ceiling truss. Cameras require an unobstructed line-of-sight to tracked rigid bodies, and participants' bodies were a major source of occlusion. I therefore dedicated extra care in setting up the cameras in a way that they captured the area around the participant's predicted location from multiple angles. I also set the origin of the world coordinate system to a point close to the center of the tracked volume to guarantee optimal calibration accuracy. See Figure 35 for a typical layout.

As mentioned above, tracked real-world objects must be augmented with an IR-reflective pattern. Throughout my studies, I used so-called rigid bodies, which are plastic clips with extruding pins for mounting IR-reflective markers. These pins allow to configure each rigid body differently, thus making them distinguishable and simultaneous tracking of multiple objects possible. The two different marker arrangements shown in Figure 36, for example, allow the software to track two separate rigid bodies. The figure also illustrates the difference in reflected lights depending on its wavelength.

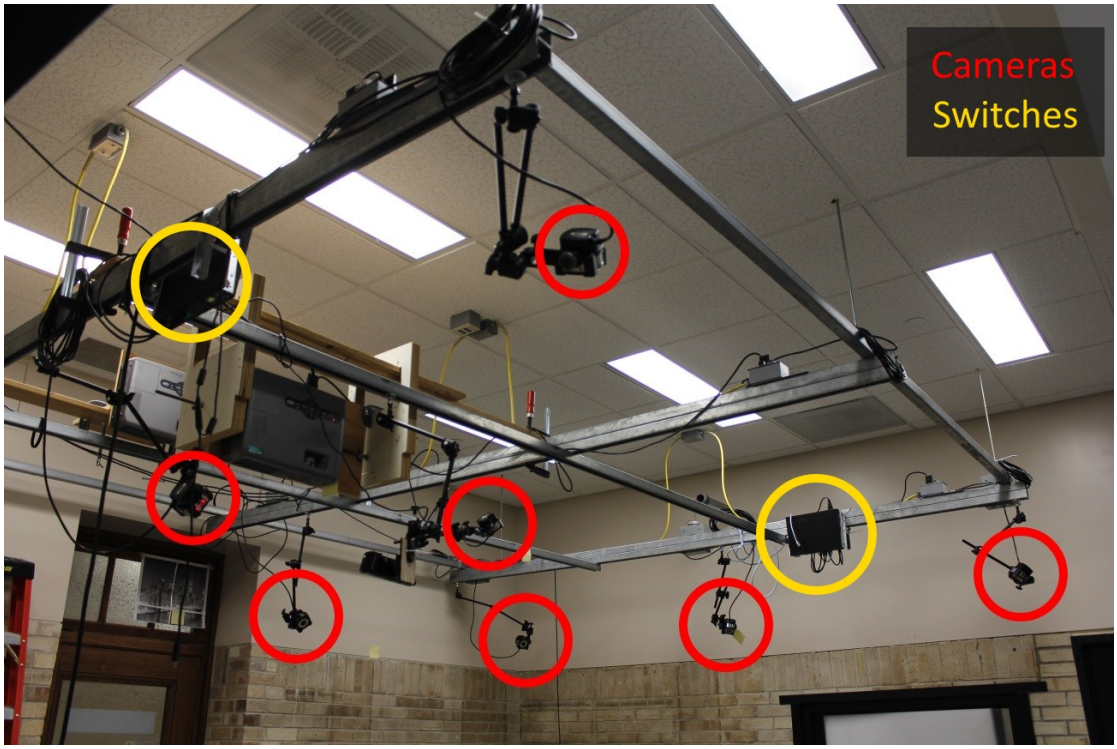



**Figure 35: Example of a seven-camera setup**

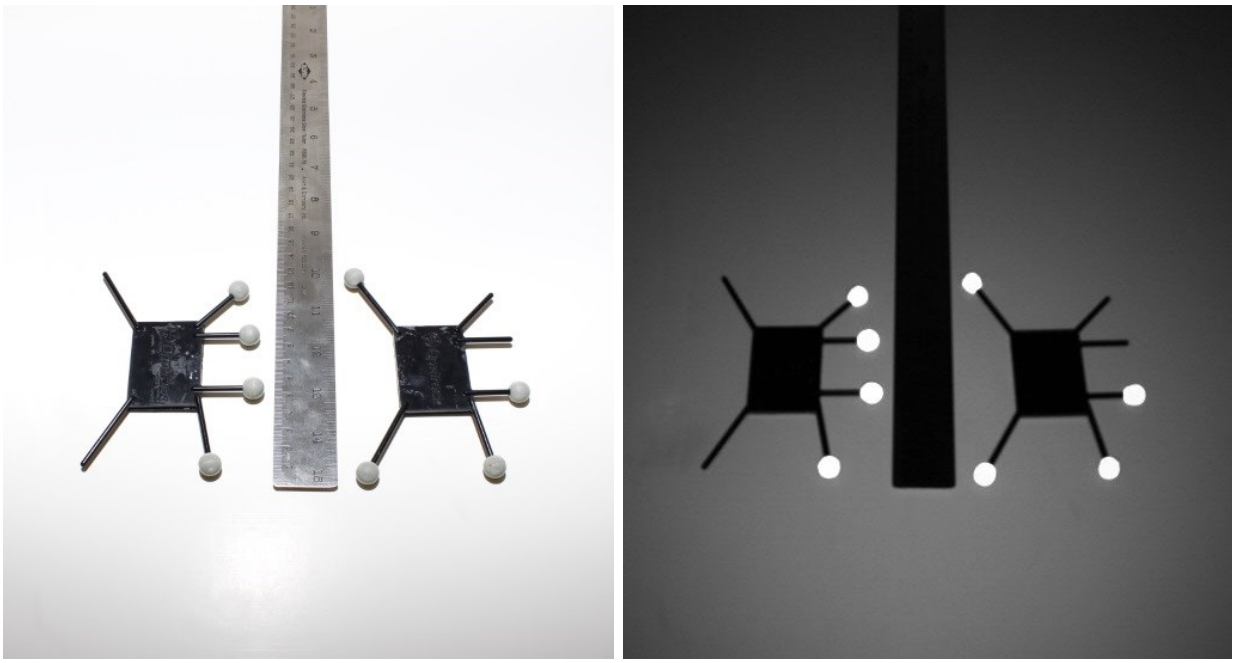


**Figure 36: Optical (left) and infra-red (right) image of two differently configured rigid bodies**

In my studies, I was mostly interested in tracking participant's pointing gestures. In order to track these gestures, I taped the rigid body to participants' extended index and middle fingers (see Figure 37).

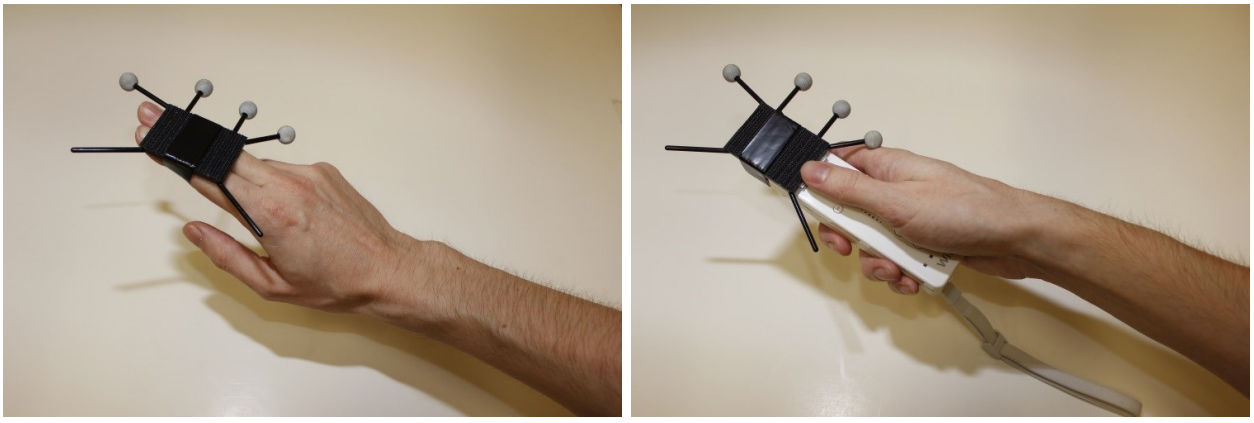


**Figure 37: Rigid body taped to a hand (left) and to a Wii Remote (right)**

## 4.2 Tracking Software, Libraries, and Custom Software

All of my studies required a mix of external libraries (e.g., for accessing tracking information) and custom-made software (e.g., user interfaces and logging). I used the Microsoft .NET Framework (version 2.0) for all of my custom software because it provides a good mix of both UI prototyping and low-level access to drivers and existing toolkits.

### 4.2.1 Tracking User Input

Polhemus provides programmatic access to their Liberty system through a Win32-library. It outputs location and orientation vectors (Tait–Bryan angles, see 4.2.4) for each sensor connected to the system unit.

Similarly, the NaturalPoint's Tracking Tools[1] allow access to tracking information via calls to its Win32-library. The tracking information for each rigid body consists of a location vector and two orientation vectors (Tait–Bryan angles and rotation quaternions). In addition, the system shipped with, a software package that provides semi-automatic camera calibration and rigid-body tracking. Figure 38 shows a screenshot for a setup with 6 cameras and 2 rigid bodies.

[1] http://www.naturalpoint.com/optitrack/products/tracking-tools/

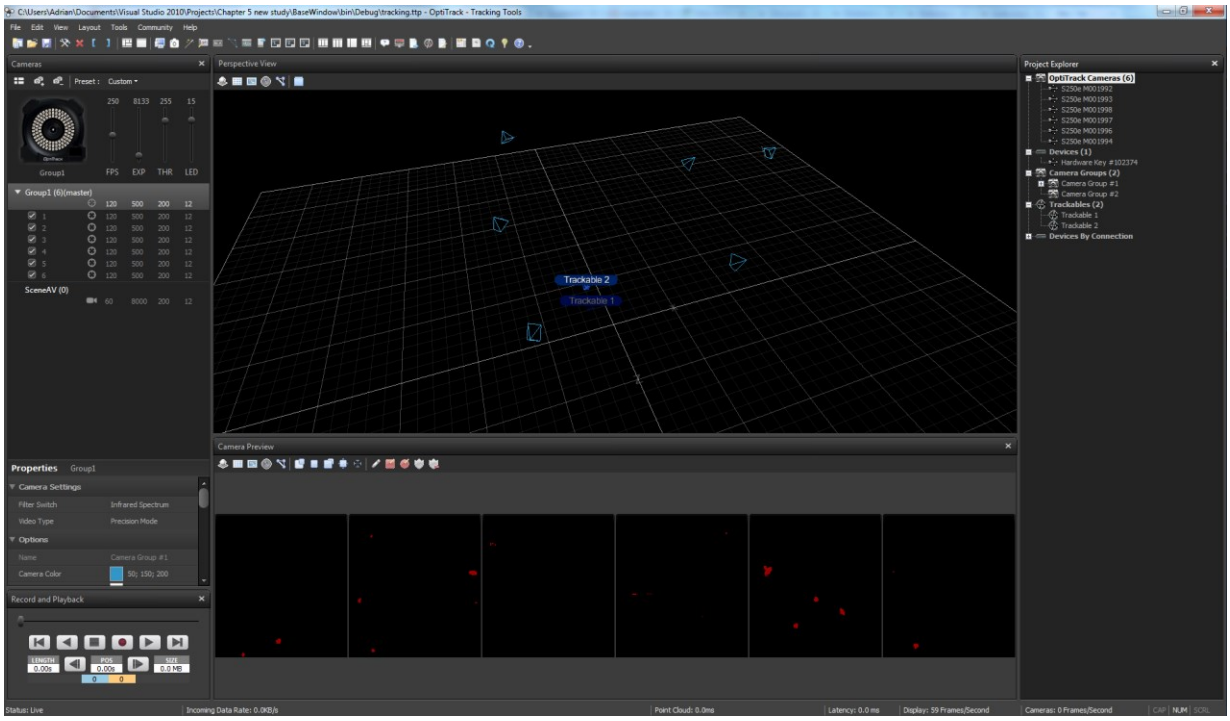



**Figure 38: NaturalPoint OptiTrack Tracking Tools**

I programmed a library (`TrackingLib`) that encapsulated the Polhemus Liberty and OptiTrack Win32-libraries in managed code (C++/CLI) and thus provided programmatic access to tracking data for all my experiment software, which was written in C#.

### 4.2.2 Mathematics Toolkit

When I started implementing *Room Pointing*, there were only few .NET-based math-libraries available, most notably AForge.NET[2]. AForge, however, did not provide all the functionality I needed for my implementation, so I decided to write my own toolkit, `FastMath`. Over the time, I kept adding functionality to `FastMath`, and as of today, it supports vector- and matrix-manipulation (e.g., arithmetics, Gaussian elimination, Gram-Schmidt orthonormalization, and QR decomposition), 1D and 2D data filtering (uniform, normal, and $\chi^2$), pseudo-random number generation (based on the Mersenne twister[3]), coordinate projection (Mercator, Mollweide, and Winkel III, see 4.2.5), and export to Excel, Matlab, and Mathematica.

[2] http://www.aforgenet.com/
[3] http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html

### 4.2.3 Modelling the Environment

There are multiple ways on how to model real-world objects. Realistic 3D-models usually use a polygon-based representation of objects. This type of modelling supports realistic collision detection, for example, by testing a vector intersecting the object's polygons. However, it is time consuming to create these 3D models, and it can be computational expensive to calculate collisions. A simpler approach uses bounding boxes or bounding spheres. These enclosing volumes are easier to create as they require substantially less information. Bounding boxes require one location vector and three orthogonal dimension vectors ($x'$-, $y'$-, and $z'$-dimensions), bounding spheres one location vector and one scalar (radius). The bounding sphere can even be further simplified by omitting the object's radius altogether, and describing the object with a single vector (the object's center).

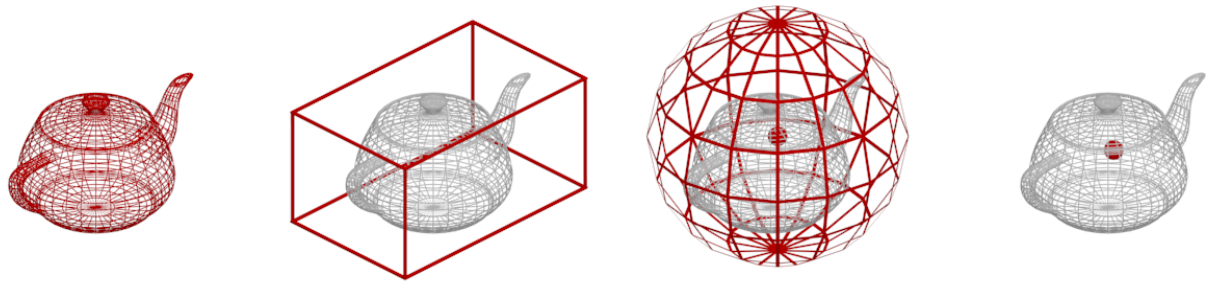


**Figure 39: Types of 3D models: realistic, bounding box, bounding sphere, and single vector (left to right)**

For my implementation, I decided to model real-world objects as single vectors with a fixed radius. There are multiple advantages to this approach. First, it makes creating models fast and uncomplicated. $x$- and $z$-coordinates (horizontal) can easily be obtained by counting floor tiles with an accuracy of up to $\sim 7.5\ cm$ (¼ feet) as floor tiles in the environment is used are exactly 1 by 1 sqft, and $y$-coordinates (vertical) can be measured with a measuring tape. Second, giving all real-world objects the same radius allows for selecting small real-world objects as selection proxies. Real-world objects inherently have different sizes, and modelling them realistically, i.e. some smaller than others, would make small objects difficult to point at accurately. When using fixed radii for every real-world object, it is also easy to adjust sizes if necessary, for example, to eliminate unassigned regions (see 4.2.5).

### 4.2.4 The Mathematics of Selecting Pointing Targets

In Euclidean space, every location $\vec{l}$ can be expressed as a 3-tuple (a vector), typically by three orthonormal dimensions $\vec{l} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$. The location is relative to the origin $\vec{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ of the overall coordinate system; coordinates relative to $\vec{0}$ are typically called *world coordinates*.

#### *Addition and Rotation*

In the context of this dissertation, two vector operations are of particular interest: addition and rotation.

Adding a vector $\vec{v}$ to a vector $\vec{l}$ by is achieved by either adding the vector components:

$$\vec{l'} = \vec{l} + \vec{v} = \begin{pmatrix} l_x \\ l_y \\ l_z \end{pmatrix} + \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = \begin{pmatrix} l_x + v_x \\ l_y + v_y \\ l_z + v_z \end{pmatrix} \quad (1)$$

or by multiplying $\vec{l}$ with an addition matrix $T_v$:

$$\vec{l'} = T_v\vec{l} = \begin{pmatrix} 1 & 0 & 0 & v_x \\ 0 & 1 & 0 & v_y \\ 0 & 0 & 1 & v_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} l_x \\ l_y \\ l_z \\ 1 \end{pmatrix} = \begin{pmatrix} l_x + v_x \\ l_y + v_y \\ l_z + v_z \\ 1 \end{pmatrix} \quad (2)$$

Any orientation in Euclidian space can be achieved by rotating a vector three times (three angles $\alpha, \beta, \gamma$ around three axes $x, y, z$). To rotate a vector $\vec{l}$, it has to be multiplied with three rotation matrixes:

$$\vec{l'} = R_x(\alpha)R_y(\beta)R_z(\gamma)\vec{l} \quad (3)$$

There are multiple notations for formalizing rotations, a particular useful one being the Tait–Bryan notation. With the Tait–Bryan notation, the three rotation angles are relative to the object coordinate system (and not the world coordinate system, see Figure 40).

$$\vec{l'} = R_{\Psi,\Theta,\Phi}l' = R_{z''}(\Phi)R_{x'}(\Theta)R_y(\Psi)\vec{l} \quad (4)$$

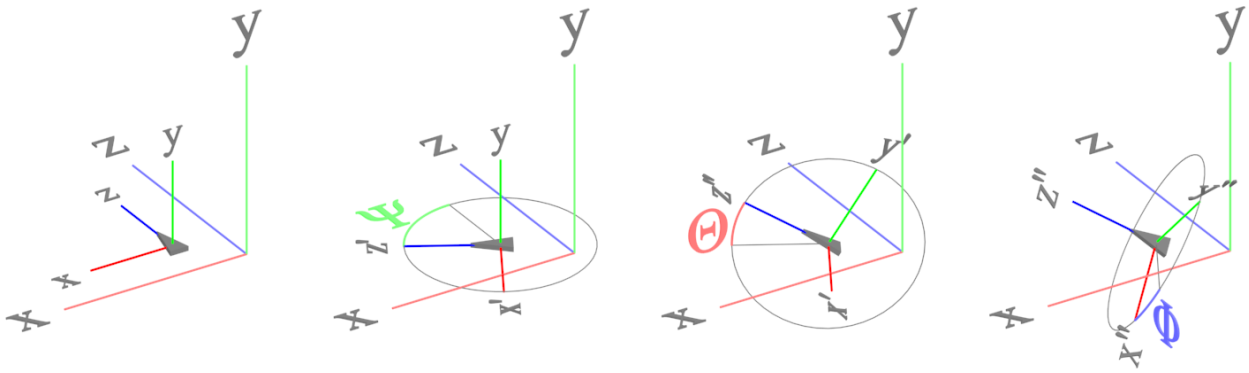



**Figure 40: Yaw-, pitch-, and roll-rotation using Tait–Bryan angles; first rotation (yaw) by Ψ around y, second rotation (pitch) by Θ around $x'$, last rotation (roll) by Φ around $z''$.**

To calculate the rotation matrix $R_{\Psi,\Theta,\Phi}$, one can either use Tait–Bryan angles $r = \begin{pmatrix} \Psi \\ \Theta \\ \Phi \end{pmatrix}$ or a rotation quaternion $q = \begin{pmatrix} q_w \\ \vdots \\ q_z \end{pmatrix}$. Since the Tait–Bryan angles returned from Tracking Tools were faulty (see 4.3.1), I decided to exclusively use rotation quaternions:

$$R_q = \begin{bmatrix} 1-2(q_y^2+q_z^2) & 2(q_xq_y-q_wq_z) & 2(q_wq_y+q_xq_z) \\ 2(q_xq_y+q_wq_z) & 1-2(q_x^2+q_z^2) & 2(q_yq_z-q_wq_x) \\ 2(q_xq_z-q_wq_y) & 2(q_wq_x+q_yq_z) & 1-2(q_x^2+q_y^2) \end{bmatrix} \quad (5)^4$$

### *Calculating Pointing Targets*

I decided to test two different methods for calculating the target of people's pointing gestures: smallest angle and shortest distance. With smallest angle, the system selects the real-world proxy with the smallest angular distance between the pointing ray and the true direction, i.e. the ray originating in people's hand and passing through the proxy. With shortest distance, the system selects the real-world proxy object with the shortest (orthogonal) distance from the pointing ray. In this chapter, I will mostly talk about the implementation; for a comparison and evaluation, see 4.3.2.

[4] http://mathworld.wolfram.com/EulerAngles.html

The pointing ray from the hand $\vec{p}$ can then be calculated by multiplying the forward vector $\vec{z} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ with the inverse rotation matrix $R^{-1}$:

$$\overrightarrow{d_p} = R^{-1}\vec{z} \tag{6}$$

In addition, one can calculate the true direction between hand and the real-world object by subtracting the hand location $\vec{l}$ from the object's location $\vec{o}$.

$$\overrightarrow{d_t} = \vec{o} - \vec{l} \tag{7}$$

## Calculating the Shortest Distance

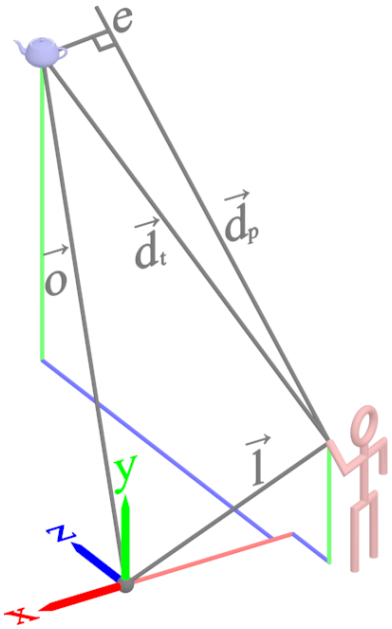



**Figure 41: Shortest distance**

For using the shortest distance to determine the pointing target, one has first to find the point $\overrightarrow{d'_p}$ on the pointing ray $\overrightarrow{d_p}$ that is closest to the real-world object $\vec{o}$. This point can be written as:

$$\overrightarrow{d'_p} = u\,\overrightarrow{d_p} \tag{8}$$

with:

$$u = \frac{\overrightarrow{d_t} \cdot \overrightarrow{d_p}}{\|\overrightarrow{d_p}\|^2} \tag{9}$$

From this, $e$ can be easily calculated:

$$e = \|\overrightarrow{d_t} - \overrightarrow{d'_p}\| = \|\overrightarrow{d_t} - u\,\overrightarrow{d_p}\| = \left\|\overrightarrow{d_t} - \frac{\overrightarrow{d_t} \cdot \overrightarrow{d_p}}{\|\overrightarrow{d_p}\|^2}\,\overrightarrow{d_p}\right\| \tag{10}$$

The final step is finding the real-world object $obj$ with the shortest distance $\varepsilon_{obj}$ within a threshold $\delta$:

$$\forall\, obj \in Obj : \min(e_{obj})\ .\, e_{obj} < \delta \tag{11}$$

The following C# code segments shows this procedure using my `FastMath` toolkit.

```
private readonly Vector PointingDirection = new Vector(3, 0.0f, 0.0f, 1.0f);
private readonly IDictionary<String, Vector> Mappings = new Dictionary<String, Vector>();

public String PointingAt(Vector Location, Vector Rotation, Single Threshold, out Single Error)
{
    Single distanceMin = Single.MaxValue;
    String result = String.Empty;

    Matrix rot = Matrix3D.RotationQuaternion(Rotation);
    Matrix rotSolved;
    GaussianEleminiation.Solve(rot, Matrix.Identity(3), out rotSolved);
    Vector dirPoint = rotSolved * PointingDirection;

    IEnumerator<KeyValuePair<String, Vector>> e = Mappings.GetEnumerator();
    while (e.MoveNext())
    {
        Vector dirTrue= e.Current.Value - Location;
        Single u = Math.Min(dirPoint * dirTrue, 0.0f);
        Single distance = (dirTrue - dirPoint * u).Norm();
        if (distance >= Threshold || distance >= distanceMin) continue;
        distanceMin = distance;
        result = e.Current.Key;
    }

    ErrorResult = distanceMin;
    return result;
}
```

## Calculating the Smallest Angle

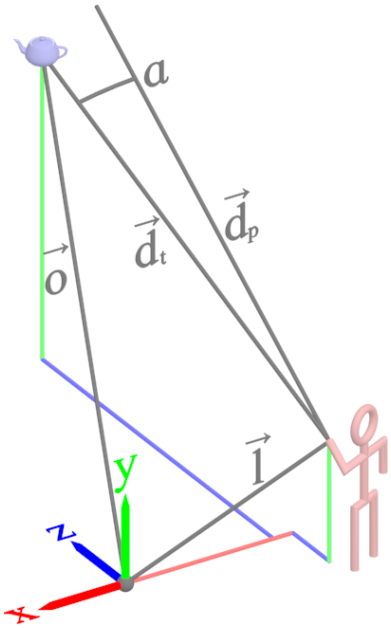



**Figure 42: Smallest angle**

For using the smallest angle to determine the pointing target, the next step after determining $\vec{d_t}$ is calculating the angle between the true direction and the pointing direction using the dot-product:

$$\alpha = \cos^{-1}\left(\frac{\vec{d_p}\cdot\vec{d_t}}{\|\vec{d_p}\|\cdot\|\vec{d_t}\|}\right) \qquad (12)$$

Finally one has to find the real-world object $obj$ for all mapped objects $Obj$ with the smallest angle $\alpha_o$ within a certain threshold $\delta$:

$$\forall\, obj \in Obj : \min(\alpha_{obj})\ .\alpha_{obj} < \delta \qquad (13)$$

The following C# code segments shows this procedure using my `FastMath` toolkit.

```
private readonly Vector PointingDirection = new Vector(3, 0.0f, 0.0f, 1.0f);
private readonly IDictionary<String, Vector> Mappings = new Dictionary<String, Vector>();

public String PointingAt(Vector Location, Vector Rotation, Single Threshold, out Single Error)
{
    Single angleMin = Single.MaxValue;
    String result = String.Empty;

    Matrix rot = Matrix3D.RotationQuaternion(Rotation);
    Matrix rotSolved;
    GaussianEleminiation.Solve(rot, Matrix.Identity(3), out rotSolved);
    Vector dirPoint = rotSolved * PointingDirection;

    IEnumerator<KeyValuePair<String, Vector>> e = Mappings.GetEnumerator();
    while (e.MoveNext())
    {
        Vector dirTrue= e.Current.Value – Location;
        Single angle = MathTools.Acos(dirPoint * dirTrue / (dirPoint.Norm() * dirTrue.Norm()));
        if (angle >= Threshold  || angle >= angleMin) continue;
        angleMin = angle;
        result = e.Current.Key;
    }
    Error = angleMin;
    return result;
}
```

I built a C# library called `RoomPointing` that implemented both aforementioned methods for calculating the pointing target using `TrackingLib` and `FastMath`. In my user studies I used these libraries to calculate the digital artifact that participants were pointing at. In addition, `RoomPointing` provided methods for reading room models from XML files and visualizing two-dimensional map data see (see Figure 43).

### 4.2.5 Visualization of Input Space and Pointing Targets

#### *Visualization and Projections of the Input Space*

Although the environment is modeled in three dimensions $(x, y, z)$, calculating the selected real-world proxy object at a given time $t$ is a two-dimensional problem since the user's hand location $\vec{l}$ is fixed at $t$: $f_{\vec{l}}(\Psi, \Theta) \Rightarrow obj$. The input space is therefore two-dimensional with $-\pi < \Psi < \pi$ and $-\frac{\pi}{2} < \Theta < \frac{\pi}{2}$. A way to conceptualize this dimension reduction is imagining the surface of a sphere with its center being $\vec{l}$ and the true directions $\overrightarrow{d_t}$ for all $obj \in Obj$ projected as points onto the surface of the sphere. One can then project the surface of the sphere onto a flat Cartesian canvas, similar to a geographic map.

A frequently used way of flattening the surface of spheres is the Mercator-projection. A problem of this projection, however, is its areal distortion: areas toward the poles ($|\Theta| \rightarrow \frac{\pi}{2}$) appear bigger than they are in reality. Typically, projections suffer from multiple types of distortion, such as

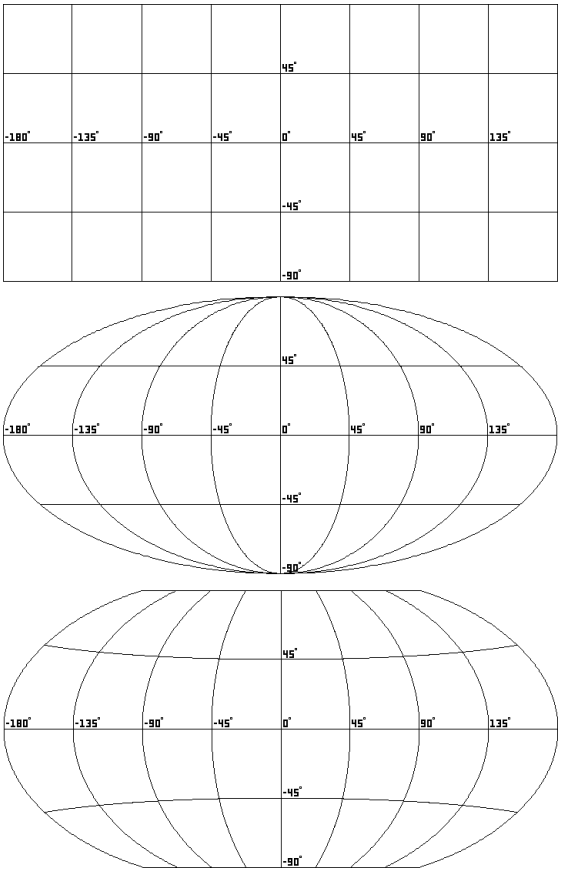



**Figure 43: Three types of map projections: Mercator (top), Mollweide (center), and Winkel-III (bottom)**

area, shape, direction, and distance. In my analyses, I was mostly interested in analyzing the influence of target area size on selection accuracy. I therefore decided to use the equal-area Mollweide-projection for visualizing the location of selection targets in the environment. To get an initial general impression of mappings I also used the Winkel-III projection, which compromises between area, direction, and distance distortions.

### *Visualization of Pointing Targets*

A good way of conceptualizing target selection is the Voronoi diagram. A Voronoi diagram is a way of partitioning a surface into regions (or cells) so that every point on the border of a region has the same distance to 2 or more seed points (Berg, Cheong, Kreveld, and Overmars, 2008). This implies that the cell's seed is the closest seed to all points within the cell, i.e. the closest seed of any point is the seed of the cell the point is in.

The formal definition of a cell $C$ is that all points $p$ within $C$ are closer to the cell's seed $s_c$ than to any other seed $s_o$:

$$\forall p \epsilon C \wedge \forall s \epsilon S \,.\, dist(p, s_c) < dist(p, s_o), c \neq o \qquad (14)^5$$

The left panel in Figure 44 shows an example of a Voronoi diagram (black) with cells (green) and seeds (red). The center panel shows the same seeds with an added constraint: the distance between point and seed cannot be larger than a threshold $\delta$:

$$\forall p \epsilon C \wedge \forall s \epsilon S \,.\, dist(p, s_c) < dist(p, s_o) \wedge dist(p, s_c) < \delta, c \neq o \qquad (15)$$

[5] Mark de Berg, Otfried Cheong, Marc van Kreveld, Mark Overmars. 2008. Computational Geometry. Springer-Verlag, Berlin / Heidelberg, Germany

One effect of this additional constraint is that it can create regions that are not assigned to any seed (gray). The value of $\delta$ hereby determines the size of these unassigned regions (see Figure 44, center and right panel).

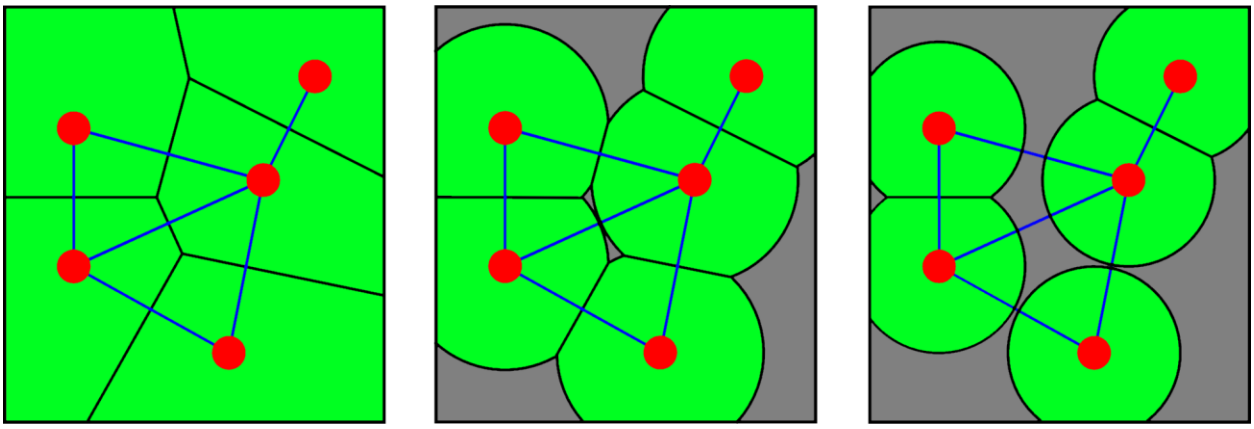


**Figure 44: Real-world objects (red) with Voronoi diagram (black), Delaunay triangulation (blue), cells (green), and unassigned regions (gray).**

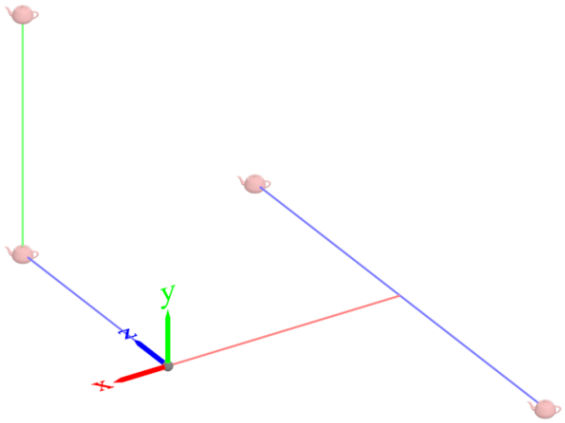



**Figure 45: Spatial layout of the Voronoi diagrams in Figure 46**

As I pointed out above, one can reduce object selection at a specific time to a $2D$ problem: $f_{\vec{l}}(\Psi, \Theta) \Rightarrow obj$. It is therefore possible to calculate a Voronoi diagram by taking the projected points (projected true directions $\overrightarrow{d_t}$) as seed points, either the angular or the Euclidean distance as the distance function $dist(q, s_i)$, and the maximum allowed pointing error as the threshold $\delta$.

Figure 46 shows two examples of such Voronoi diagrams in Mollweide projection. Both examples were generated using my `RoomPointing` library. The room model consisted of four real-world objects located at $\begin{pmatrix}0\\0\\1\end{pmatrix}$, $\begin{pmatrix}-1\\0\\1\end{pmatrix}$, $\begin{pmatrix}0\\1\\1\end{pmatrix}$, and $\begin{pmatrix}-1\\0\\-1\end{pmatrix}$, the location $\vec{l}$ is fixed at $\begin{pmatrix}0\\0\\0\end{pmatrix}$ (see Figure 45). From $\begin{pmatrix}0\\0\\0\end{pmatrix}$, the four real-

world objects are projected to $\binom{0°}{0°}$, $\binom{0°}{45°}$, $\binom{45°}{0°}$, and $\binom{0°}{135°}$ (red points), where they act as seeds for the Voronoi diagram. The left example in Figure 46 shows the Voronoi diagram as well as the differently colored cells for sufficiently large $\delta$ to prevent unassigned regions ($\delta > \frac{5}{8}\pi$). In the right example, $\delta$ is small enough ($\delta = \frac{1}{3}\pi$) to allow for unassigned regions (gray).

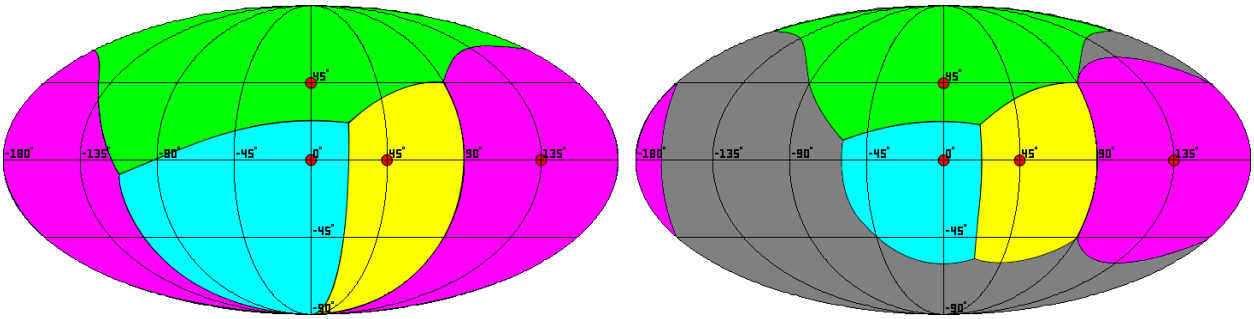



**Figure 46: Voronoi diagram in Mollweide projection; red points are the projection of the four targets from Figure 45; cells are colored, unsigned regions remain gray.**

I use these visualizations throughout my dissertation to visualize location and density of real-world proxy objects; they were not shown to participants.

### 4.2.6 Technology-related Terminology in Room Pointing

Now that I have established the technological background of my implementation for room-based interaction, I want to describe and illustrate some terms that I am using throughout my dissertation, in particular real-world proxy-object, pointing target, and target zone.

Real-world proxy object: dotted area (1). A real-world object that acts as a selection proxy for a digital artifact. In this example, the real-world object is “top of the red ladder”.

Pointing target: green dot (2). In my implementation, the pointing target of a real-world proxy object (1) is modeled by a single vector. This vector also acts as the seed of the target zone (3).

Target zone: green area (3). Pointing within the target zone of the real-world object (1) selects the digital artifact that is associated with the selection proxy (1). Mathematically, the target area is the Voronoi-cell that corresponds to the seed (2). The cell is located on an imaginary sphere that has its center at the origin of the pointing ray (4).

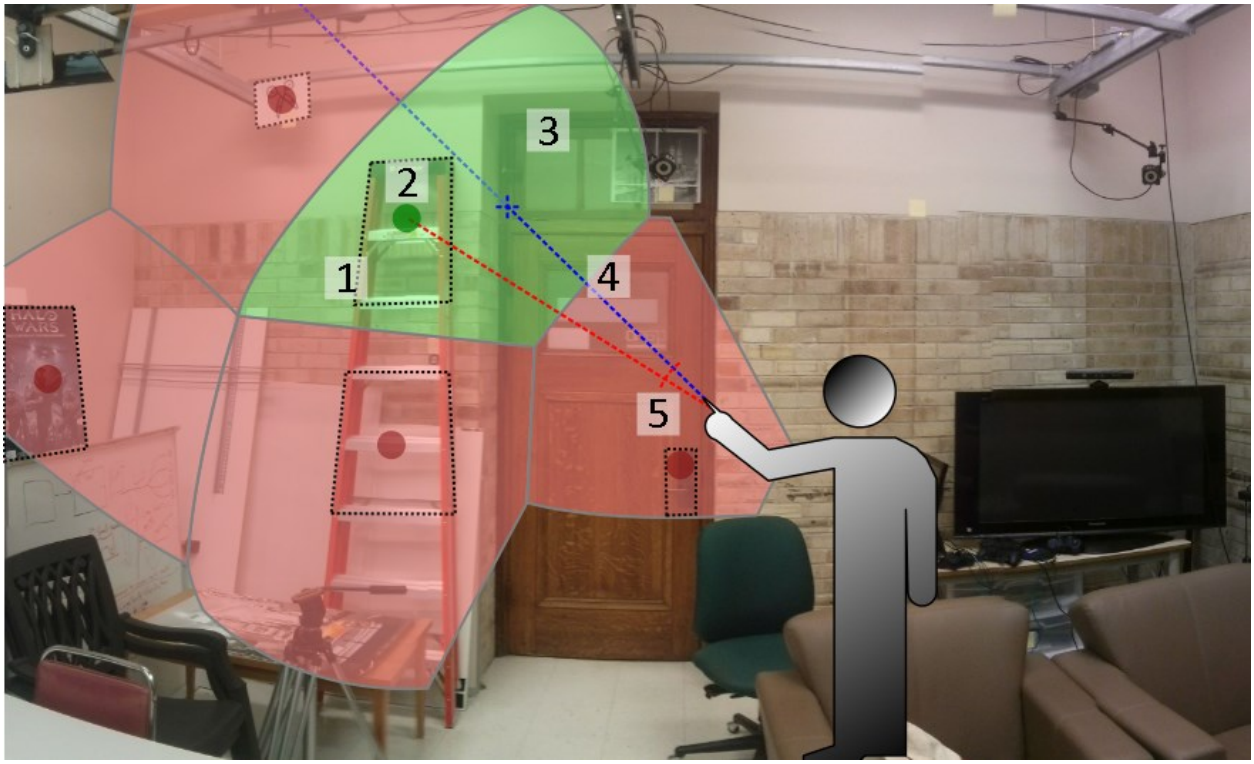



**Figure 47: Example and terminology of room-based interaction**

Pointing ray: blue dotted line (4). This ray represents the direction and origin of the mid-air full-arm pointing gesture. The blue cross represents the intersection of the pointing ray with a target zone or, mathematically, a cell on the imaginary Voronoi-sphere (3).

Pointing error: red dotted arc (5). In my implementation, the error is the angular difference between the pointing ray (4) and an imaginary line from the origin of the pointing ray the model-representation of the proxy object (2).

### 4.2.7 System Overview

Overall, I implemented three libraries that served as the foundation for all my experiment software. `TrackingLib` provided unified access to the two proprietary hardware systems from a managed .NET environment, `RoomPointing` was used to calculate the selected real-world proxies from people's pointing gestures, and `FastMath` provided the necessary linear algebra functionality. Figure 48 shows how these three libraries (blue) interact with each other as well as

with the proprietary software drivers (red) and experiment software used in Chapters 5 through 7 (green).

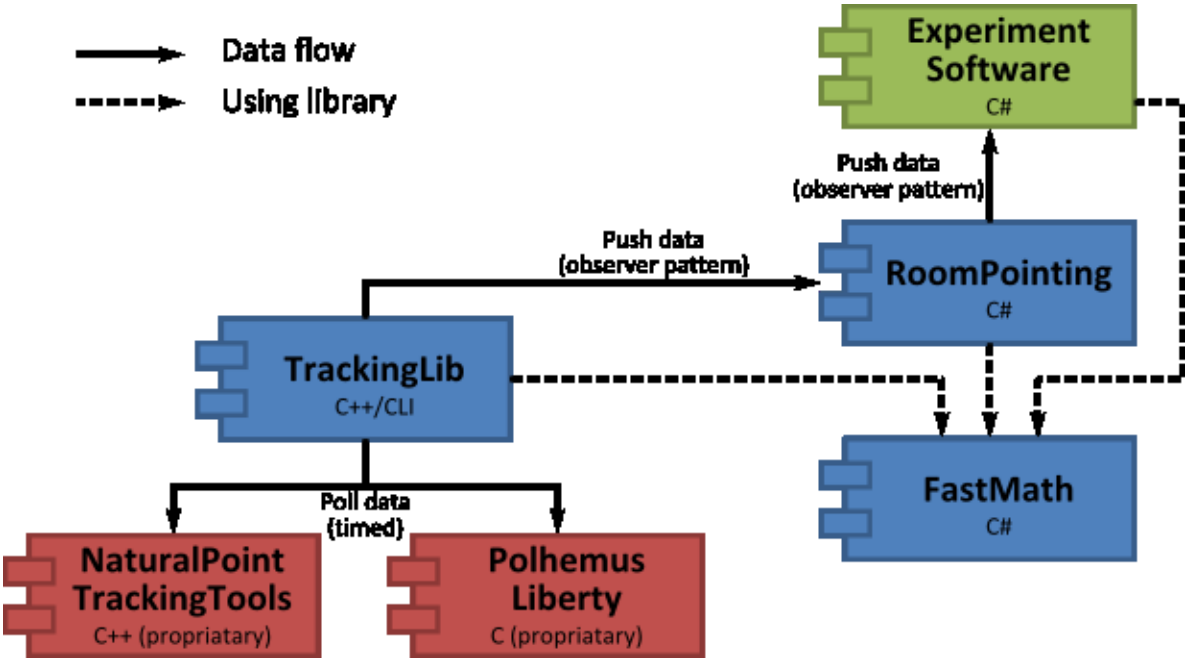



**Figure 48: UML diagram of the libraries used in this dissertation; existing libraries in red, my own libraries in blue, experiment software in green**

## 4.3 System Evaluation

### 4.3.1 Hardware and Software

I conducted an informal evaluation of both the Polhemus Liberty tracker and the NaturalPoint OptiTrack system. Although both systems have different strengths and weaknesses, it became clear that an optical tracking system would be better suited for the purpose of my studies than an electromagnetic. The main reason is that the Liberty tracker's orientation accuracy decreased within the first 1 $m$ to a degree that the pointing direction could not be calculated accurately enough, i.e. tracking accuracy became worse than people's full-arm pointing capabilities. OptiTrack's driver and software package, on the other hand, was less stable and contained some severe bugs that made development and debugging difficult. First, location- and orientation-vectors follow different Cartesian notation conventions: locations are assigned a left, rotations a right handedness. This problem can be solved easily by negating values from one dimension.

Second, Tait–Bryan angles are not independent. This means that, for example, changing yaw (Ψ) also changes pitch (Θ) and roll (Φ). At first, it appeared that this behavior was due to not rotating the rigid bodies coordinate system, i.e. the second rotation appear to be relative to the world coordinate system and not to the object coordinate system. While this dependency could have been easily corrected, a more thorough investigation revealed that there were other factors linking Θ and Ψ. Thankfully, the rotation quaternions were correct, and I was able to use them for input into my `TrackingLib`.

The general-purpose PCs that I used in my user studies provided sufficient computational power to for my software. `TrackingLib` has an overall complexity level of $O(m \times n)$, where $m$ is the number of tracked people and $n$ is the number of modelled real-world proxy-objects. This limits the scalability of my system to some degree, i.e. when tracking thousands of people and objects, but does not affect the system within the scope of its intended use case (domestic environments, fewer than 10 tracked people, and fewer than 100 modelled real-world proxy objects).

Overall, current hardware is capable of accurately tracking people's location, their arm movement, and the orientation of their fingers, and existing software allows access to all necessary data. Software libraries for processing and interpreting tracking data, however, still has to be written from scratch. Since this is only a minor hurdle, I felt confident that I would be able to capture and analyze people's mid-air full-arm pointing gestures accurately enough for conducting further research.

### 4.3.2 Algorithm for Calculating Pointing Targets

I also conducted an informal evaluation of both using pointing angle ($\delta$) and pointing distance ($\varepsilon$) for calculating pointing targets. Both error metrics are mathematically linked as $\tan\delta = {\varepsilon}/{|\vec{d}|}$, where $\vec{d}$ is the vector from the user to the orthonormal projection of the pointing target onto the pointing direction vector $\overrightarrow{d_p}$. Figure 49 illustrates that with pointing distance as error metric (top), the system would select the closer target 2 over the more distant target 1 ($\varepsilon_2 < \varepsilon_1$), and with pointing angle as error metric (bottom), the system would select the more distant target 1 over the closer target 2 ($\delta_1 < \delta_2$). In general, using pointing distance favors closer objects

when calculating pointing targets, using pointing angle favors distant objects. Mathematically, the equation $\tan\delta = {}^{\varepsilon}/_{|\vec{d}|}$ explains this effect:

$$\varepsilon_2 = {}^{\tan\delta}/_{|\overrightarrow{d_2}|} < \varepsilon_1 = {}^{\tan\delta}/_{|\overrightarrow{d_1}|} \; . \; |\overrightarrow{d_2}| < |\overrightarrow{d_1}| \text{ and}$$

$$\delta_1 = \tan^{-1}{}^{\varepsilon}/_{|\overrightarrow{d_1}|} < \delta_2 = \tan^{-1}{}^{\varepsilon}/_{|\overrightarrow{d_2}|} \; . \; |\overrightarrow{d_2}| < |\overrightarrow{d_1}|$$

This effect, however, disappears when all pointing targets have similar distances from the user: $|\overrightarrow{d_1}| \approx |\overrightarrow{d_2}|$. Since this was the case in all of my experiments and no real-world proxy objects were substantially closer to the user than others, I assumed that the algorithm would not influence the people's performance and even perception of the selection technique.

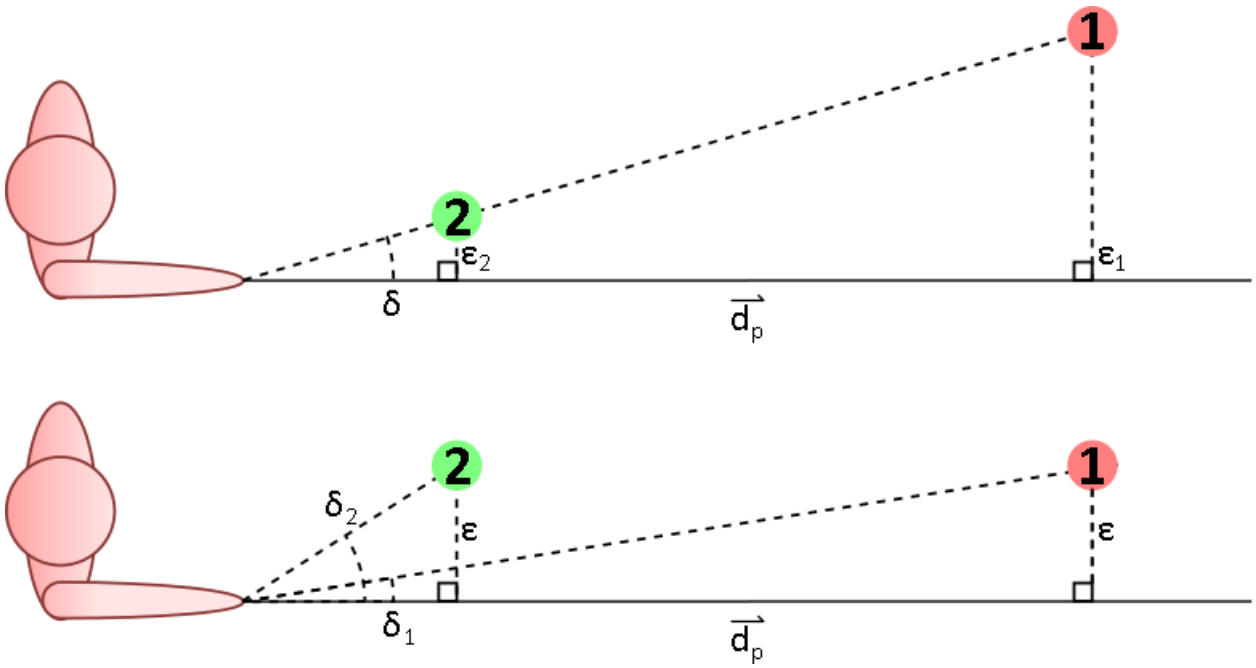



**Figure 49: Comparison of pointing angle (error term $\delta$) and pointing distance (error term $\varepsilon$) as measurement for calculating pointing targets**

Since, to my knowledge, all existing research uses pointing direction as error metric, I decided to do the same in order to make my results more comparable with the ones existing literature.

# Chapter 5 Performance of Room-based and Menu-Based Selection Interfaces

In Chapter 1, I argued that current interaction techniques for Human-Environment Interaction, i.e. in-place or navigation-based interfaces, can be slow, inconvenient, disruptive, or physically and mentally demanding to use. In this chapter, I set out to validate this claim and show that pointing-based interaction can be a viable alternative to touch-based interaction. For this, I will compare navigation-based with pointing-based interaction techniques. Overall, I am interested in answering two research questions:

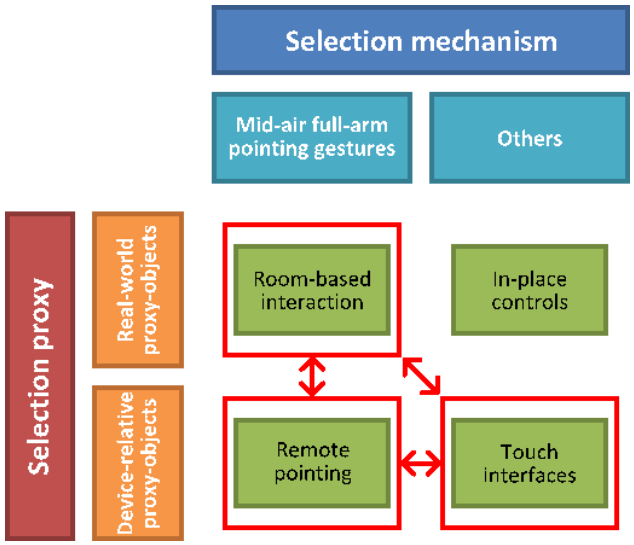



Figure 50: Design space of study 1

1. To what degree does the type of selection mechanism influence peoples' performance?
2. To what degree does the type of selection proxy influence peoples' performance?

## 5.1 Interacting with Smart Environments

Smart Environments offer people many opportunities for interaction that are different from traditional desktop interfaces. These opportunities oftentimes occur when people are trying to complete a goal outside of traditional desktop computing, such as checking the weather forecast before leaving home, or outside of the digital realm altogether, such as turning on the lights in the living room. As I laid out in Chapter 1, today these interactions mostly occur through in-place controllers, such as light switches and on-device buttons, or through navigation-based interfaces, such as interfaces on smart phones and tablets.

In-place controllers have the critical disadvantage in that users have to walk up to them. This process alone makes in-place interfaces an order of magnitude slower than any interaction people can perform *in-situ*. For this reason, I exclude in-place interfaces from this performance-based experiment.

To mitigate this disadvantage, numerous researchers suggested the use of smart phones (Ballagas et al., 2006; Barkhuus and Polichar, 2011; Rukzio et al., 2006) as people tend to keep them within reach. These now-ubiquitous devices use navigation-based interfaces (WIMP) for human-computer interaction where direct touch input replaced the traditional indirect mouse input. In navigation-based interfaces, selection proxies take the form of on-screen icons. The limited amount of screen real-estate poses, however, a problem on these hand-held devices. There are two common strategies to solve this issue: arranging icons in scrollable lists (scrolling icon storage space) or shrinking icons to make more of them fit on a single screen (flat icon storage space). Both approaches have advantages and disadvantages. Lists can store a vast number of selection proxies, but they are slower because users have to navigate through the list to find the correct icon. Shrinking selection proxies avoids this kind of time-consuming navigation, but the reduced icon size makes icons more difficult to select (Fitts' Law, see 2.2.3). Both designs exist in today's touch-screen user interfaces. Scrollable lists of icons are used, for example, in the Netflix UI (iOS, Android, Web-based) where movies that are arranged horizontally by genre. Flat design is used in systems such as Google Android, where people can adjust the grid-size of their home screen to fit more icons on a single screen.

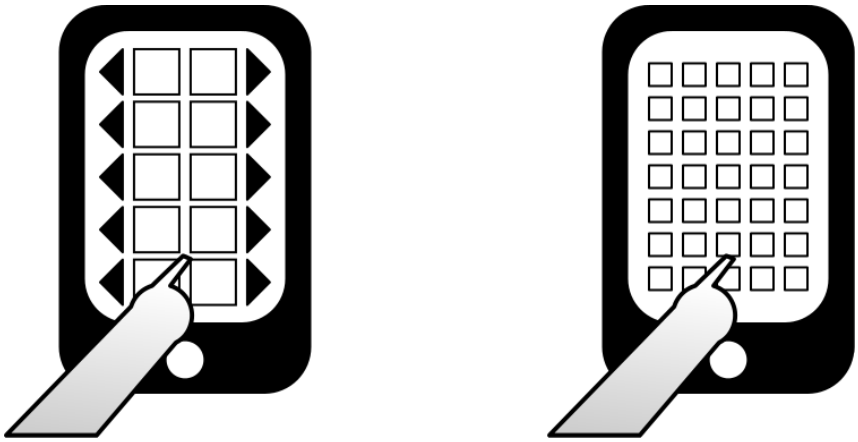


**Figure 51: Navigation-based interfaces: scrollable list (left) and flat design (right)**

As an alternative, other researchers investigated the use of large screens for fast and device-free system interaction, such as, the *Gyro Point* and the *Remote Point* (MacKenzie and Jusoh, 2001). For these interaction techniques, people make selections from proxy items (e.g., icons) displayed

on a TV screen by moving an on-screen cursor through arm- and hand-motions. The Nintendo Wii Remote and the Microsoft Kinect are commercially available examples of such an interaction technique. This type of interaction does not require users to hold a physical input device, so they can interact with the smart environment without having immediate access to their smart phones. However, the interaction still requires the presence of a device—the screen—in the environment.

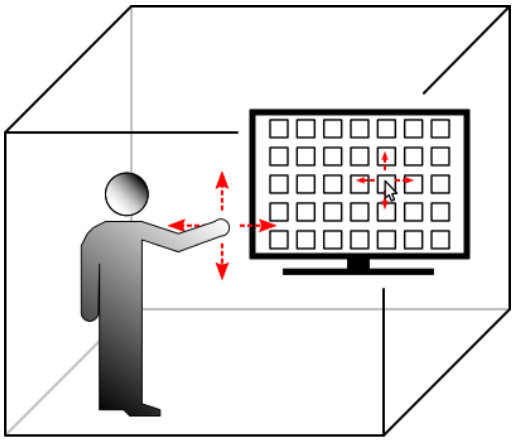


**Figure 52: Pointing-based interface using on-screen selection proxies**

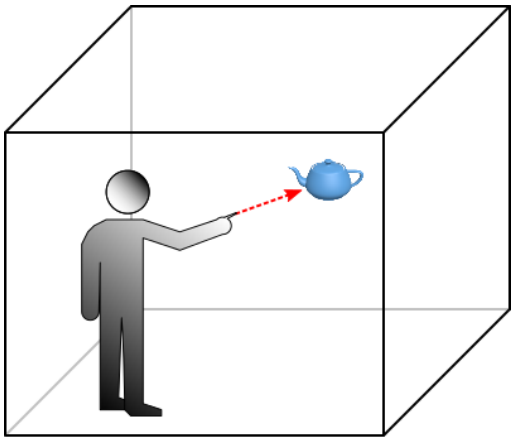


**Figure 53: *Room Pointing*: pointing-based interface using real-world proxy objects**

A second alternative, which does not require any device for displaying a UI or any visual system feedback, is using room-based interaction (*Room Pointing*). In this variation of full-arm pointing,

people point at a real-world object to select the associated system functionality or digital artifact. This type of interaction requires neither a hand-held input devices nor devices for system feedback.

Assuming touch screens to be the default standard for interaction with smart environments, the questions remains whether the two alternative pointing-based techniques can match touch interaction in terms of selection accuracy and selection time. If so, it would show them to be viable alternatives to touch input with the additional benefit of allowing device-free and feedback-free interaction.

To answer this question, I conducted a user study. I implemented four different types of selection techniques for smart environments. In *Touch Scroll*, people made selection by tapping selection proxies (icons) displayed on a hand-held device (see Figure 54.1). People had to scroll left and right to access all digital artifacts. In *Touch Flat*, all proxies were displayed concurrently, although with smaller icons (see Figure 54.2). In *Screen Pointing*, selection proxies (icons) were displayed on a TV screen, and people could make selections by moving around an on-screen cursor through full-arm pointing gestures and clicking on icons; this is similar to commercially available products like the Nintendo Wii (see Figure 54.3). Finally, in *Room Pointing* real-world objects acted as selection proxies, and people could select commands by making a full-arm pointing gesture toward the real-world object that was associated to the command (Figure 54.4). I also tested how people's performance changed in all four techniques after adding additional items, which required more scrolling in *Touch Scroll*, more precise button taps and clicks in *Touch Flat* and *Screen Pointing*, and more precise pointing in *Room Pointing*.

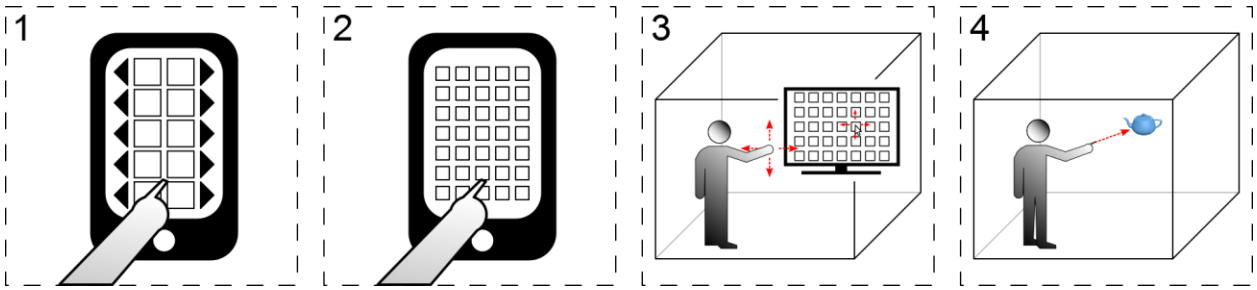



**Figure 54: *Touch Scroll* (1), *Touch Flat* (2), *Screen Pointing* (3), and *Room Pointing* (4)**

I looked at four issues:

- Organization of storage space: does flattening the input space reduce selection time without sacrificing accuracy? (*Touch Scroll* versus *Touch Flat*)
- Selection mechanism: to what degree does the type of selection mechanism (pointing-based versus touch-based) influence selection speed? (*Touch Flat* versus *Screen Pointing*)
- Proxy type: to what degree does the type of selection proxy influence selection time, accuracy, and learnability? (*Screen Pointing* versus *Room Pointing*)
- Overall selection speed and accuracy: can *Screen Pointing* and *Room Pointing* be considered viable alternatives to touch-based interaction? (*Screen Pointing* and *Room Pointing* versus *Touch Flat*)

From the previous analyses in Chapter 3 as well as existing research, I formulated three hypotheses:

1. People can make selections faster and with the same accuracy using a flat storage space compared to a linear storage space.
2. People can use pointing-based interaction with the same levels of accuracy and speed than touch-based interaction.
3. People can use room-based selection proxies with the same level of speed than screen-based selection proxies; given the differences in feedback modes, room-based interaction might show lower accuracy than screen-based interaction.

## 5.2 Study Conditions

The following sections provide details on the implementation of the four study techniques.

### 5.2.1 Touch Screen (Touch Scroll and Touch Flat)

My implementations for touch screens replicated a user interface typically found in today's smart phones, phablets, and tablets. A typical procedure for making a selection using a one of these devices would be to first unlock the screen to access the home menu, then start the UbiComp interaction app, and finally find and select the appropriate icon to tap on. Figure 57 (left) demonstrates how a typical home screen might look like; the bottom-right icon starts the UbiComp control app while the other icons remain blank to minimize distraction. As with many

mobile devices, the lack of screen real-estate poses a problem when the number of selection options exceeds the available space for proxy icons. *Touch Scroll* is designed after the aforementioned scrollable list design, which can be found, for example, in the Netflix UI for iOS and Android. Figure 55 shows a close-up of the interface where only three of six TV shows are visible at the same time. In order to select other shows, users have to click on the right arrow to scroll the menu. Scroll button are $1.0 \times 2.5\ cm^2$ of size, selection buttons $2.3 \times 2.5\ cm^2$ (width $\times$ height).

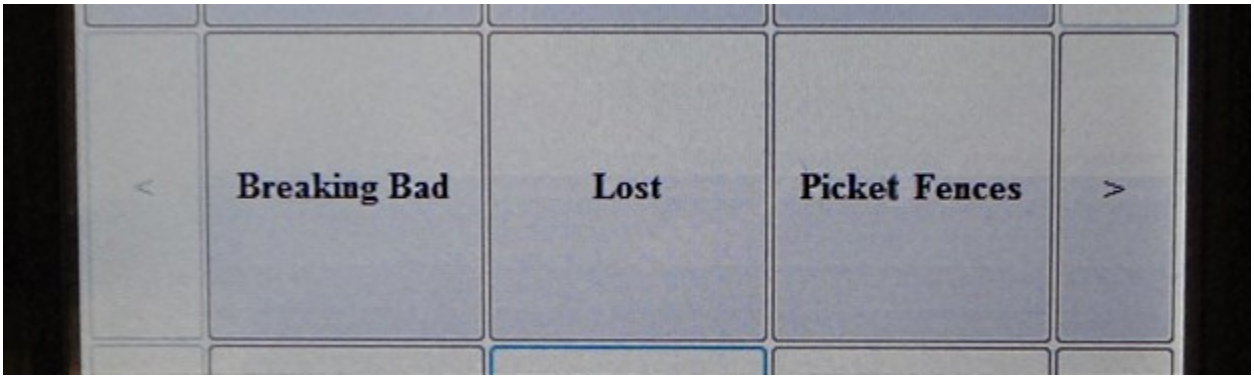



**Figure 55: Example of horizontal scrolling in *Touch Scroll***

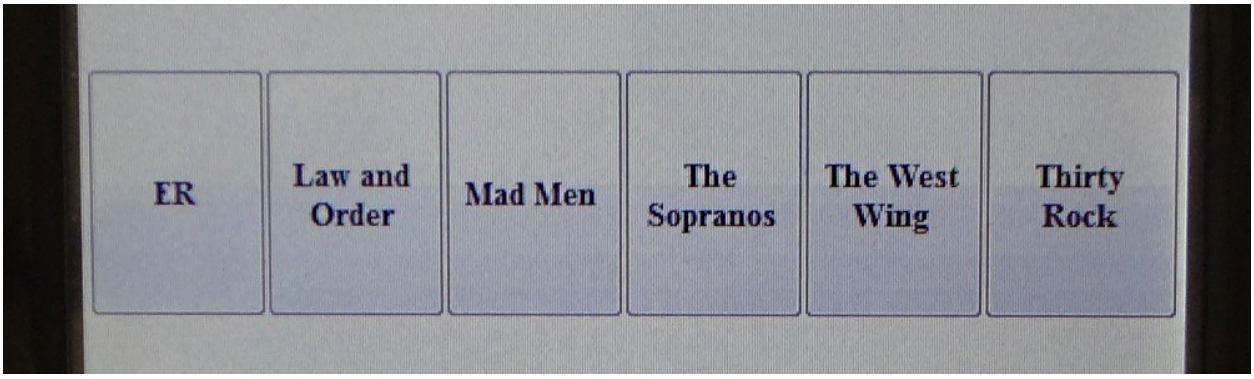



**Figure 56: Example for flat design in *Touch Flat***

*Touch Flat* is designed after the aforementioned flat design, which can be found, for example, on the Google Android home screen. Here my solution is to decrease the size of the buttons, so that all six of them fit in one row. On one hand, the reduced size makes buttons more difficult to hit,

and on the other hand, the permanent availability of all buttons renders scrolling unnecessary and thus reduces selection time (see Figure 56). With $1.4 \times 2.0\ cm^2$ (width × height) selection buttons in *Touch Flat* have half the area of the ones in *Touch Scroll*.

Figure 57 shows an overview of the touch screen interface. The home screen (left) is the first screen participants encounter. From there, they have to tap on "Room Control" to open up the UbiComp control window. Depending on the condition, this window either shows all selectable items in scrollable lists (*Touch Scroll*, Figure 57 center) or with compressed buttons (*Scroll Flat*, Figure 57 right)

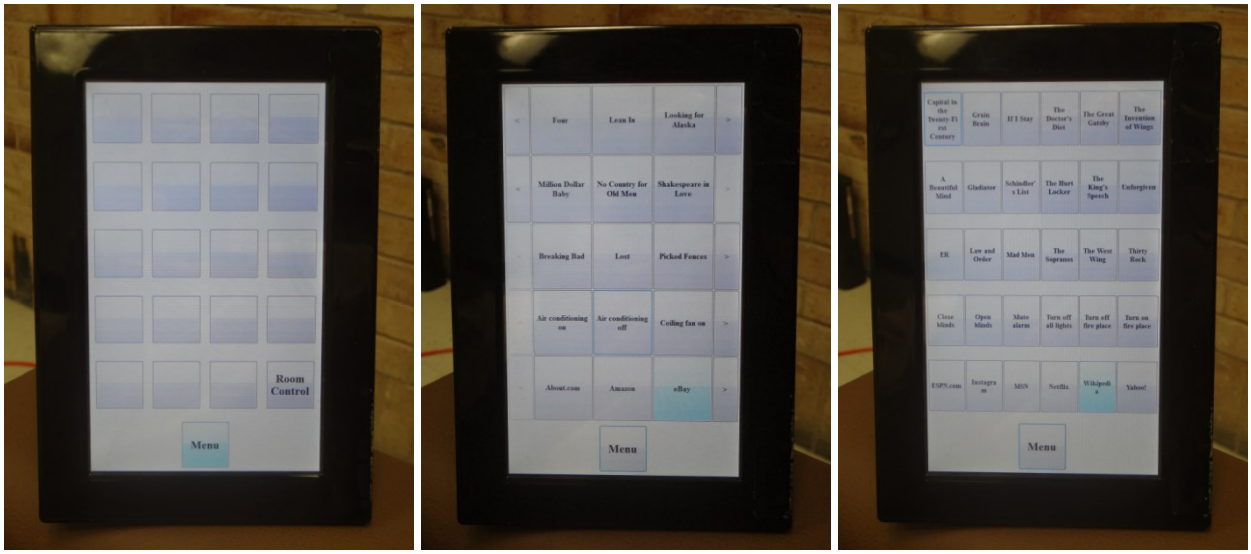



**Figure 57: Home screen (left), *Touch Scroll* (center), and *Touch Flat* (right)**

### 5.2.2 Screen Pointing

*Screen Pointing* follows the classic window / pointer UI design and has been implemented in commercial products, such as the Nintendo Wii Remote and the Microsoft Kinect. People select digital artifacts by moving an on-screen cursor over an icon (selection proxy) and confirming the selection. The main differences to touch-screen interactions is that input is now indirect. There are two major questions to answer when designing for full-arm pointing at screen-based proxies.

First, a designer has to decide which of the 6 input dimensions $(x, y, z, \varphi, \theta, \psi)$ to use. One possibility is using pointing direction $(\varphi, \theta)$, where yaw and pitch are mapped to on-screen $x$-

and $y$-coordinates (e.g., Nintendo Wii); another is using location $(x, y)$, where $x$- and $y$-coordinates of, for example, the user's hand is mapped to on-screen $x$- and $y$-coordinates (e.g., Microsoft Kinect). I decided to use pointing direction because pointing is widely used in human communication to refer to out-of-reach objects (Ekman and Friesen, 1981) and it allowed a direct comparison to *Room Pointing*.

Second a designer has to decide on a $^{c}/_{d}$-ratio between limb and cursor movement. There are multiple types of $^{c}/_{d}$-ratios (e.g. below, equal, or larger than one; constant, linear, or higher-level functions; and time- or location-dependent) (Blanch et al., 2004); their choice depends on factors like user age or expertise (Smith et al., 1999). For this study, I expected participants to have little experience with full-arm pointing interaction techniques. Subsequently, I minimized the $^{c}/_{d}$-ratio by allowing the maximal input range of $\pm 16°$ of for radial / ulnar wrist deviation (cursor: up / down) and $\pm 40.0°$ for wrist flexion / extension (horizontal cursor movement). In general, these values follow the recommendations from existing literature (Liskowsky and Seitz, 2014). I decided, however, to use an extension value below the possible maximum ($62°$) to accommodate both left- and right-handed users. This means that each icon has a size of $\sim 6.5°$ horizontally and $\sim 5.0°$ vertically in physical input space. In participant's visual field, each icon is $\sim 3.4°$ (vertical) times $\sim 2.6°$ (horizontal).

### 5.2.3 Room Pointing

One factor that sets *Room Pointing* apart from *Screen Pointing* is the difference in selection proxy design: instead of on-screen icons, *Room Pointing* uses real-world objects as proxies. This changes the procedure on how people select digital artifacts. Instead of remembering the navigation path to an artifact and visually searching for it on screen, people have to remember the association between artifact and real-world proxy, remember the location of the real-world proxy in the environment, and perform a pointing gesture toward it. Although it might appear that this procedure requires high cognitive load, human memory is specialized to perform all three steps exceptionally well (see Chapter 3).

In *Room Pointing*, real-world objects fulfill a dual role of being selection proxies as well as mnemonic devices. The richness of meaning that people can associate with real-world objects (e.g., their shape, color, location, history, name) increases their utility as mnemonic device and

can help people remember the association between real-world proxy and digital artifact better (see 2.5.5).

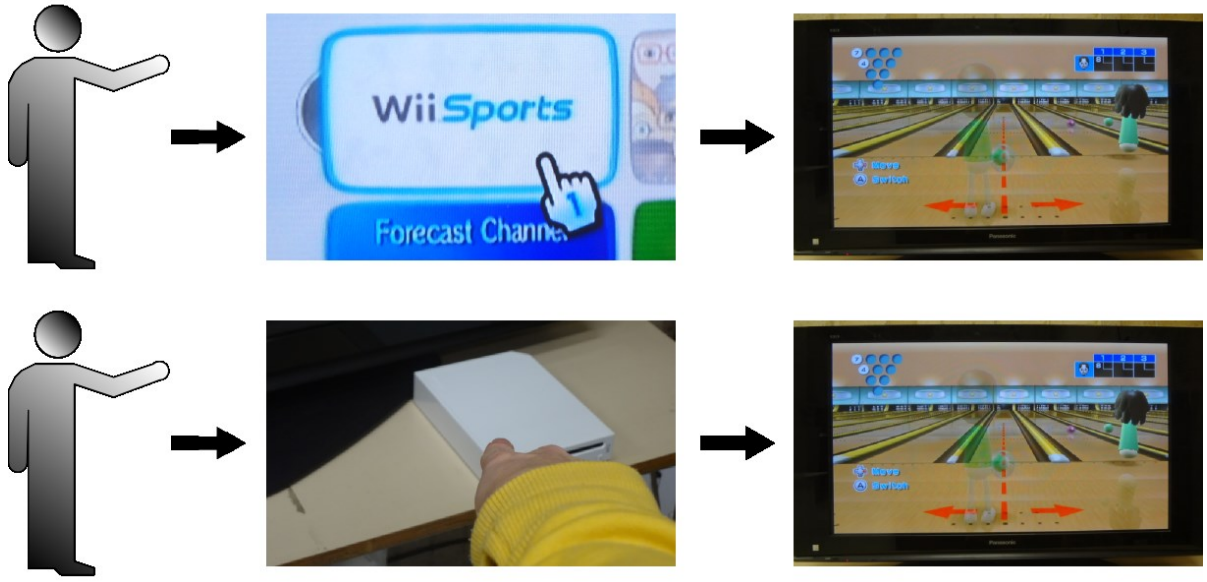



**Figure 58: *Screen Pointing* (top) and *Room Pointing* (bottom)**

## 5.3 Experimental Setup

### 5.3.1 Study Design, Participants, and Apparatus

The study used a single-factor within-participant design with interaction technique as four-level factor (*Touch Scroll*, *Touch Flat*, *Screen Pointing*, and *Room Pointing*). The order of appearance was balanced using a Latin square.

I recruited 16 participants (4 female, 12 male; ages $21 - 36$, $\bar{x} = 27$ years; 3 left-, 13 right-handed) from a local university. All participants had experience with traditional computer systems and owned a smart phone; six participants have previously used full-arm gesture control. They received a $10 honorarium for participating in this one-hour-long study.

The study was carried out in a laboratory, in which I recreated a living-room-like setting with a couch, a 42″ TV screen ($2.1\ m$ to the couch), and a mobile 7″ touch screen with $5{:}3$ aspect ratio

and 133 $dpi$ resolution (see Figure 59, left) that was connect to my experiment computer via USB. I chose this distance so that the field of view for the TV and the touch screen were comparable. For both pointing-based techniques, I tracked participant's gestures using an OptiTrack infrared tracker. I set its sampling rate to 40 $Hz$ and used a Butterworth-filter to remove frequencies above 12.5 $Hz$ to remove the effects of hand jitter.

### 5.3.2 Adding Digital Artifacts to the Environment

In this study, I also wanted to simulate a behavior I deemed typical in domestic smart environments: adding more digital artifacts. One can imagine, for example, adding more cooking recipes or e-books to one's repository. I therefore added 10 digital artifacts in the final two blocks of each condition (trails$^+$).

### 5.3.3 Study Conditions and Procedures

I implemented four different interaction techniques, two touch- and two pointing-based. Both *Touch Scroll* and *Touch Flat* mimicked current smart-phone like interaction. In *Touch Scroll*, participants could only see 15 buttons at a time, so they had to scroll left or right if the desired button was currently not displayed. In *Touch Flat*, all 30 (40 in trials$^+$) buttons were shown concurrently, although at half the size (area) than in *Touch Scroll*. The two pointing-based techniques were *Screen Pointing* and *Room Pointing*.

After filling out a consent form and an initial questionnaire, participants were seated on the couch. In both touch-based conditions, participants were handed the touch-sensitive screen (see Figure 59, left); in both pointing conditions, the touch screen was placed on a stool in front of the couch, and the participants were handed a tracked Wii Remote (see Figure 59, right). I used the trigger button ("B") on the Wii Remote for selection confirmation and calculated participants' pointing direction using the rigid bodies taped to the Wii Remote (*Screen Pointing*) or to the participant's index and middle finger (*Room Pointing*). Every $1-2$ $s$ (randomized, uniformly distributed), a pop-up on the touch screen asked participants to select a given digital artifact. There were a total of 30 (40 in trials$^+$) artifacts, divided in five categories: books, movies, TV series, environment commands, and bookmarks. Each selection technique had a different set of artifacts to avoid learning effects across conditions; I picked artifacts from known sources (Academy and Emmy Award winners, Alexa Ranking, B&N bestseller list) to make the sets comparable. At the beginning of the experiment, the system randomly selected 7 of the 30

possible artifacts. These were the 7 artifacts participants were asked to select during the experiment. In my analysis, I considered seven selections (each artifact once in random order) as one block.


**Figure 59: Touch interface (left) and pointing controllers (right)**

Each participant went through three phases (practice, trials, trials$^+$) and performed a total of $28 + 56 + 14 = 98$ selections per technique (or $4 + 8 + 2 = 14$ blocks). Generally, the practice and the trials phases were identical and just separated by a pop-up window that gave the experimenter a chance to check in with the participants; for *Room Pointing*, however, I turned off the continuous feedback about participant's current pointing target (see Figure 60, top right) with the beginning of the trials phase. This made *Room Pointing* a system-feedback-free technique. In the trials$^+$ phase, 10 new artifacts were added to the user interface, resulting in more scrolling (*Touch Scroll*) , more and smaller on-screen buttons (*Touch Flat*: new button size $1.1 \times 2.0\ cm^2$ ; and *Screen Pointing*), and smaller target zones (*Room Pointing*, see Figure 64).

In both *Touch Scroll* and *Touch Flat*, participants had to click a home button on the touch-screen, which brought them back to the main menu. After this, they had to start the "Room Control" app by tapping on a prominently located icon on the main screen. Finally, they had to find the correct item to select on the screen (see Figure 60, top left and top center).

In the *Screen Pointing* condition, participants could use their arm to move an on-screen cursor and a click with the B-button on the Wii Remote to confirm their selection. For the selection, participants had to click once with the Wii Remote to bring up the "Room Control"-menu and then select the artifact from the menu (see Figure 60, bottom).


**Figure 60: *Touch Scroll* (top left), *Touch Flat* (top center), *Room Pointing* (top, right), and *Screen Pointing* (bottom)**

In the *Room Pointing* condition, participants had to point at the correct real-world objects and confirm the selection using the Wii Remote's B-button. While pointing during the practice-

phase, participants received feedback about the current selection and the associated real-world proxy (see Figure 60, top right).

After each interaction technique, participants filled out a NASA TLX form; after the experiment, which lasted for 1 hour, they were paid a $10 honorarium.

### 5.3.4 Data Analysis

As an initial step, I removed all $3\sigma$-outliers from the time-data in order to account for unusual participant behavior, such as playing around with the system; I calculated $\sigma$ per participant and per phase (practice, trials, trials$^{+}$). This amounted 3.2 % of all data to be removed, which is higher than expected (99.7 %) and could indicate that my data was not perfectly normal-distributed.

For RM-ANOVAs, I used Greenhouse-Geisser correction for non-spherical data and Bonferroni correction for post-hoc tests.

## 5.4 Results

### 5.4.1 Completion Time

A regression analysis over blocks 1 – 12 (practice and trials) revealed that completion time was logarithmically dependent on block number, as predicted by the power law of practice (*Touch Scroll*: $R^2_{adj} = .45, \beta = -0.9$; *Touch Flat*: $R^2_{adj} = .45, \beta = -0.9$; *Screen Pointing*: $R^2_{adj} = .33, \beta = -1.1$; *Room Pointing*: $R^2_{adj} = .39, \beta = -1.0$; see Figure 61).

#### *Performance with the initial set of artifacts*

A 4 × 12 RM-ANOVA with technique and block (#1 – #12) as factors indicated significant main effects for both factors ($F(2.0,30.0) = 88.9, p < .001$ and $F(1.5,22.9) = 34.7, p < .001$; Mauchly's test: both $p < .05$) and a significant interaction ($F(2.2,32.5) = 3.2, p < .05$). A post-hoc analysis revealed that *Screen Pointing* was significantly faster and *Touch Scroll* significantly slower than the other three techniques (all $p < .001$). Block #5 was the last block that was significantly slower than any of the following blocks (#10 – #12, all $p < .05$); when comparing block #6 with all following blocks (#7 – #12), I could not detect any significant improvements in completion time any more (all $p > .05$); this transition came one block after the switch from

practice to trials. For this reason, I assumed that participants reached their best performance in blocks #6 – #12, and thus used them as ground truth for the comparison with the trials$^{+}$ phase.

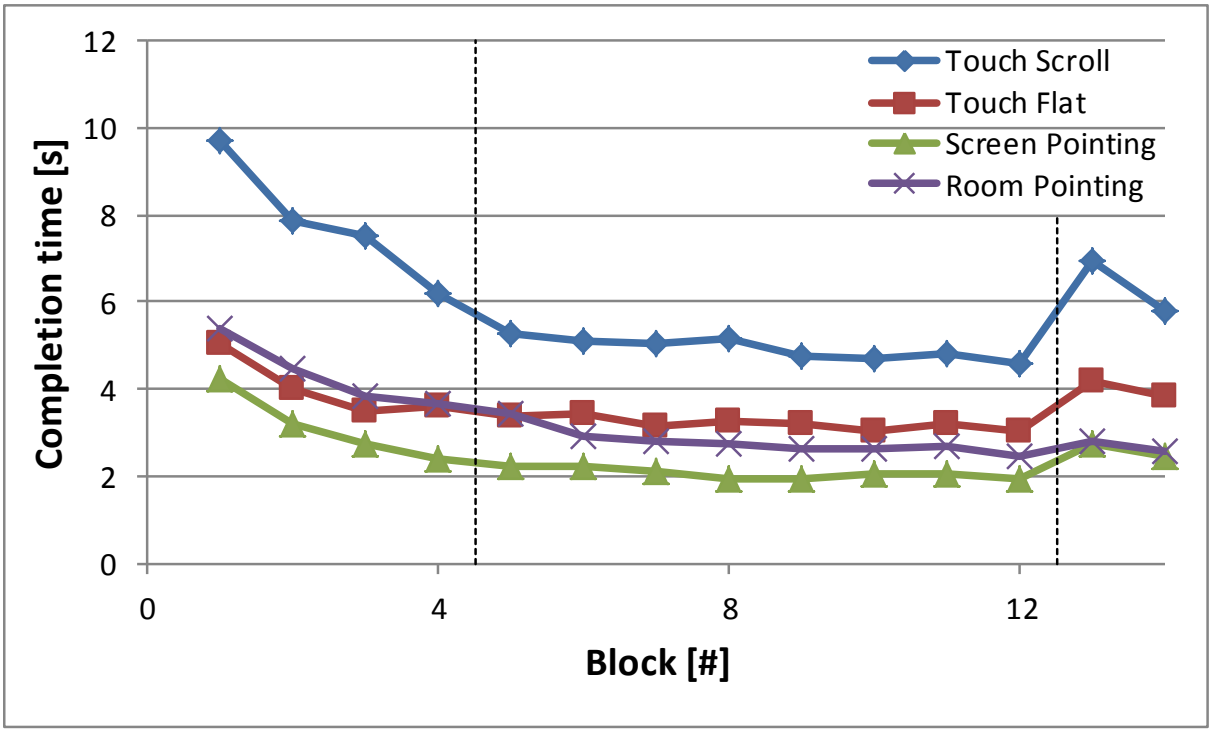



**Figure 61: Completion times; the vertical lines mark the transitions between practice, trials, and trials$^{+}$**

**Table 4: Mean completion time and standard error**

| | Practice | Trials | Trials$^{+}$ |
|---|---|---|---|
| **Touch Scroll** | $7.5\,s \pm 0.27\,s$ | $5.0\,s \pm 0.07\,s$ | $6.2\,s \pm 0.22\,s$ |
| **Touch Flat** | $4.0\,s \pm 0.13\,s$ | $3.2\,s \pm 0.03\,s$ | $3.9\,s \pm 0.10\,s$ |
| **Screen Pointing** | $3.1\,s \pm 0.19\,s$ | $2.0\,s \pm 0.02\,s$ | $2.3\,s \pm 0.07\,s$ |
| **Room Pointing** | $4.2\,s \pm 0.17\,s$ | $2.7\,s \pm 0.05\,s$ | $2.6\,s \pm 0.11\,s$ |

### *Performance after adding artifacts*

A $4 \times 9$ RM-ANOVA with technique and block (#6 – #14) as factors indicated significant main effects for both factors ($F(3,45) = 112.2, p < .001$ and $F(4.0,60.5) = 18.9, p < .001$; Mauchly's test: $p < .01$ for block) and a significant interaction ($F(24,360) = 4.7, p < .001$).

Completion time was significantly higher in block #13 than in the previous seven blocks (#6 – #12). Unlike *Room Pointing*, *Touch Scroll*, *Touch Flat*, and *Screen Pointing* showed a significant increase in selection time after adding 10 items between blocks #12 and #13 (all $p < .05$).

### 5.4.2 Selection Accuracy

#### *Performance with the initial set of artifacts*

A 4 × 12 RM-ANOVA with block (practice and trials phase) and technique as factors indicated a significant main effect for technique ($F(3,45) = 6.6, p < .05$) and a significant interaction ($F(33,395) = 2.0, p < .05$). A post-hoc analysis showed that *Room Pointing* produced significantly more errors than *Touch Flat* and *Screen Pointing* (both $p < .05$).

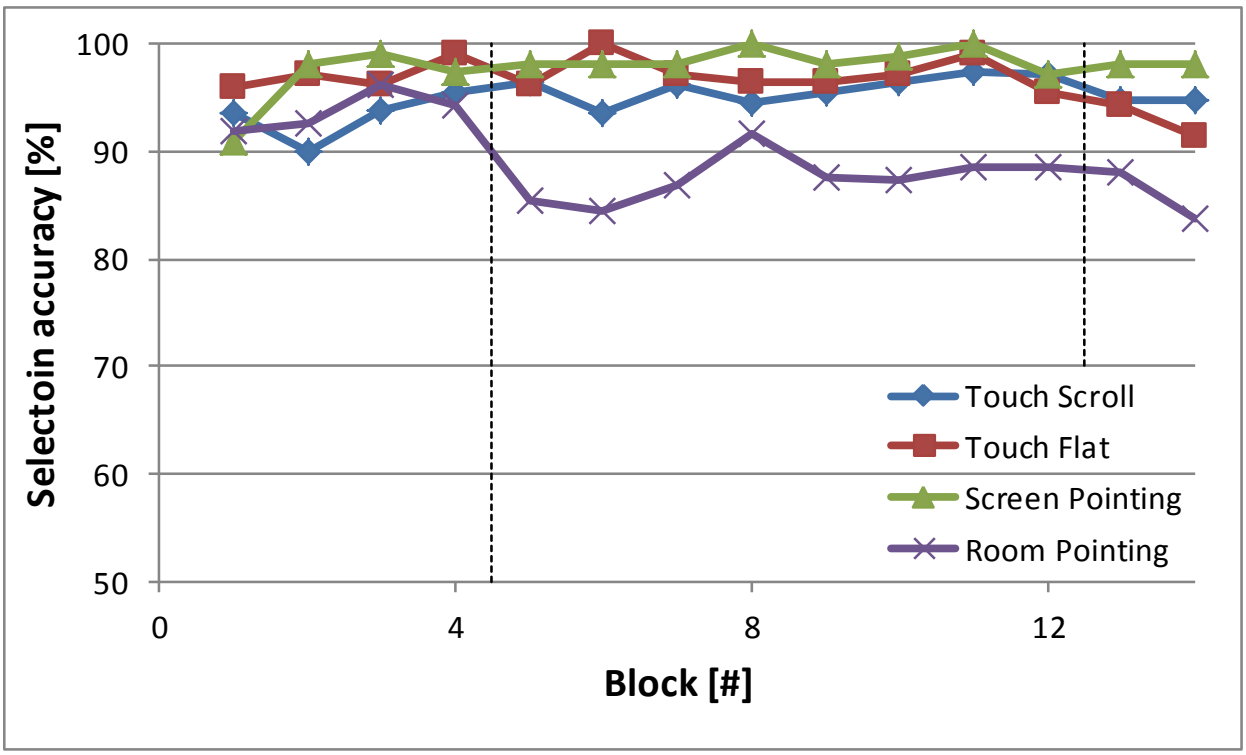



**Figure 62: Selection accuracies;**
**vertical lines mark the transitions between practice, trials, and trials$^+$**

*Room Pointing* showed a 14 % accuracy drop after removing real-time feedback about the participant's current pointing target (from block #4 to #5). While this drop was not significant, block #5 was significantly less accurate than blocks #2 and #3 (both $p < .05$).

### *Performance after adding artifacts*

A 4×9 RM-ANOVA with technique and block (#6 – #14) as factors indicated only a significant main effect for technique ($F(3,45) = 10.3$, $p < .001$) and no interactions. As before, this was due to the lower accuracy of *Room Pointing* compared to the other three interaction techniques (all $p < .05$). The transition between trials and trials$^+$ showed no significant effect on any technique.

**Table 5: Mean selection accuracy and standard error**

| | Practice | Trials | Trials$^+$ |
|---|---|---|---|
| **Touch Scroll** | 94 % ± 1 % | 96 % ± 1 % | 95 % ± 2 % |
| **Touch Flat** | 97 % ± 1 % | 98 % ± 1 % | 93 % ± 1 % |
| **Screen Pointing** | 96 % ± 1 % | 98 % ± 0 % | 99 % ± 1 % |
| **Room Pointing** | 96 % ± 1 % | 88 % ± 1 % | 85 % ± 2% |

### 5.4.3 Demographics and Task Load Index

I asked participants to rate from 1 – 5 (never – constantly) how often they carry their smart phones at their body when being at home. I assumed that most participants would carry their phones half of the time (3). A $\chi^2$-test, however, revealed that participants tend to carry phones significantly more often than expected ($\bar{x} = 3.4$; $\sigma = 1.31$; $\chi^2(4, N = 16) = 13.1$, $p = .01$).

I analyzed the TLX using a $4 \times 6$ RM-ANOVA with technique and rating as factors indicated only one significant main effect for technique ($F(3,51) = 4.8$, $p < .01$) and no interaction. A post-hoc analysis revealed that participants rated *Screen Pointing* significantly lower (better) than *Room Pointing* and *Touch Flat* significantly lower than *Touch Scroll* (both $p < .05$).

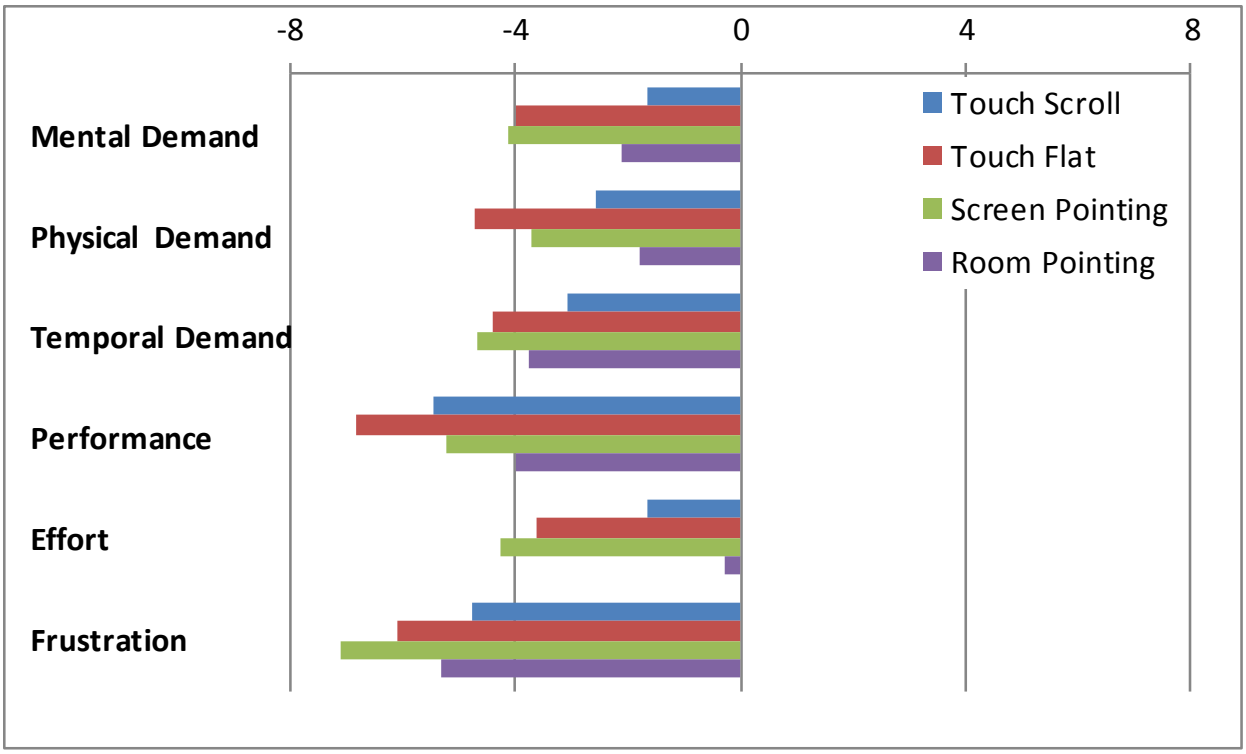



**Figure 63: TLX scores (from −8 to +8; lower scores are better)**

## 5.5 Discussion

In the discussion, I first review the three hypotheses that I set out to verify with this experiment. I then describe additional findings that emerged from the results. Finally, I address the limitations of this research.

### 5.5.1 Review of the Main Hypotheses

At the beginning of this chapter, I formulated three hypotheses about the expected results from this study:

1. Flat storage space is better than hierarchical storage space.
2. Performance with a pointing-based selection mechanism is as good as touch-based selection mechanism.
3. Using real-world interaction proxies has no disadvantage over a screen-based interaction proxies in terms of selection speed, but possibly in terms of selection accuracy.

### *Storage Space*

The results of the experiment show that using a flat input space significantly reduces completion time throughout all blocks in trials and trials$^+$ (*Touch Scroll* vs. *Touch Flat*: all $p < .01$). At the same time, I could not find a significant difference in selection accuracy ($p > .1$). This confirms that, at least for icons with at least 1 $cm$ width, a flat input space allows for faster selection without sacrificing selection accuracy.

### *Selection Mechanism*

The results show that pointing techniques (particularly *Screen Pointing*) can perform as well as, or better than, touch-based techniques. In contrast to touch interaction, *Screen Pointing* allows people to interact with smart environments in a device-free manner. I demonstrated that this advantage is not offset by lower selection time or accuracy. In contrary, my results indicate that *Screen Pointing* was significantly faster than *Touch Flat* throughout all blocks (all $p < .001$) while showing no significant differences in accuracy ($p > .1$). While this result might seem to be surprising, the actual difference in selection time (apx.$1\ s$) is minimal in the context of this research, and it may be too small to favor one technique over the other. The difference could be an artifact of my implementation—in particular, participants required one additional button click (opening the main menu) in *Touch Flat* over *Screen Pointing*. However, even after subtracting this additional step, *Screen Pointing* is at least as fast as touch-based input. (See 5.5.4 for an in-depth discussion on this issue.) Overall, I conclude that both touching on and pointing toward a screen perform equally well in the context of making selections in smart environments.

### *Selection Proxy Type*

With *Room Pointing*, people gain the additional ability to perform feedback-free interaction. Just like with *Screen Pointing*, participants showed very similar selection time compared to the "default" standard touch interfaces. The added feature of feedback-free interaction comes without a selection time penalty. I found, however, a significant difference in selection accuracy during some, but not all blocks in trials and trials$^+$ between *Room Pointing* and *Screen Pointing*. I will discuss this in more detail below. Overall, I found that the difference in proxy type between *Screen Pointing* and *Room Pointing* had a significant impact on selection accuracy.

I also confirmed the advantage of spatial stability in the user interface. Unlike *Room Pointing*, both touch-based techniques and *Screen Pointing* showed a small but significant increase in selection time after adding 10 digital artifacts.

#### *Selection Speed and Accuracy*

The results from this study show that people can use pointing-based interaction at least as accurately and quickly as touch-based interaction as long as people have feedback about which selection proxy they have currently selected. This means that people can use device-free interaction without any penalty. Without that feedback, people can still make selections as quickly, though at reduced selection accuracy. In summary, I feel confident in recommending pointing-based over touch-based interaction for the numerous scenarios in smart interaction where device- and feedback-free interaction is beneficial. My study showed a clear trade-off for Room Pointing between selection accuracy and the ability to perform feedback-free interaction, i.e. a 10 % drop in accuracy for 30 mapped digital artifacts.

### 5.5.2 Device-Free Interaction

In the introduction (see 1.2 and 5.1), I advertised *Room Pointing* and *Screen Pointing* to be device-free interaction techniques (i.e., users do not have to hold a device in their hands); in my study, however, I chose to let participants hold physical input devices (a Wii Remote and a rigid tracking body). The reason for this is that current device-free tracking devices do not provide the technological capabilities to track human pointing gestures accurately enough; by using them, I would have measured device limitations, not human capabilities. I am, however, confident that my study results are generalizable. For *Room Pointing*, the rigid body neither affected the user's mental model, the kinematics of the arm and hand movement, nor the afferent feedback channels involved in producing the pointing gesture. It only slightly inhibited free index and middle finger movement due to the size of the rigid body. This is also mostly true for *Screen Pointing*. The only differences are the finger and, arguably, the wrist posture.

One aspect I did not address is segmentation ambiguity or more commonly called: Midas Touch (Morris et al., 2010). This is the phenomenon that a system might have difficulties distinguishing between (intentional) commands issued by the user and (unintentional) deictic gestures or emblems. There are multiple existing methods for solving this problem, such as using dwelling

to indicate the completion of a gesture or selection or simple sign gestures to indicate the beginning.

### 5.5.3 Room Pointing and the Effect of Adding Digital Artifacts

One distinct feature of *Room Pointing* is that the spatial layout does not change when adding more mappings between real-world proxies and digital artifacts to the environment. Instead of shifting, target zones around pointing targets simply shrink in size. Previous research has shown that this spatial stability benefits expert users (Fitts and Posner, 1967). Figure 64 (top) shows the layout and distribution of the 30 real-world proxies in my study from the participant's viewpoint. (To make target areas comparable, I used Mollweide-projection, an equal-area map projection, for my visualization.) Figure 64 (bottom) shows the effects of adding 10 more targets (new pointing targets colored in yellow, pointing zones in grey). It is apparent that users do not have to re-learn existing selections, they just have to be more precise when producing pointing gestures.

Selection accuracy for *Room Pointing* was substantially lower than with the other techniques, and I see four possible explanations for this finding—in addition to the fact that *Room Pointing* is a memory-based technique rather than a feedback-based technique. First, people might simply not recall the association between digital artifact and real-world proxy object correctly, thus point toward the wrong selection proxy. Second, I observed that participants were too reliant on system feedback during the practice phase: instead of watching their arm movement as they would do for distal pointing, participants focused on observing the system feedback on the screen. When I turned off system feedback after block #4, selection accuracy dropped approximately 10%. After this, participants changed their strategy in order to regain satisfactory accuracy levels, a behavior corroborated by the following increase in accuracy. Third, there appears to be a higher variance in selection accuracy for different people for *Room Pointing* ($\bar{x} = 0.87$, $\sigma = 0.09$, $min = 0.70$, $max = 0.98$) compared to *Screen Pointing* ($\bar{x} = 0.99$, $\sigma = 0.02$, $min = 0.95$, $max = 1.0$). This indicates that some people were able to achieve equally high performance in both techniques. Fourth, the TLX indicated that the higher physical demand and effort could have led to fatigue during the 98 *Room Pointing* trials.

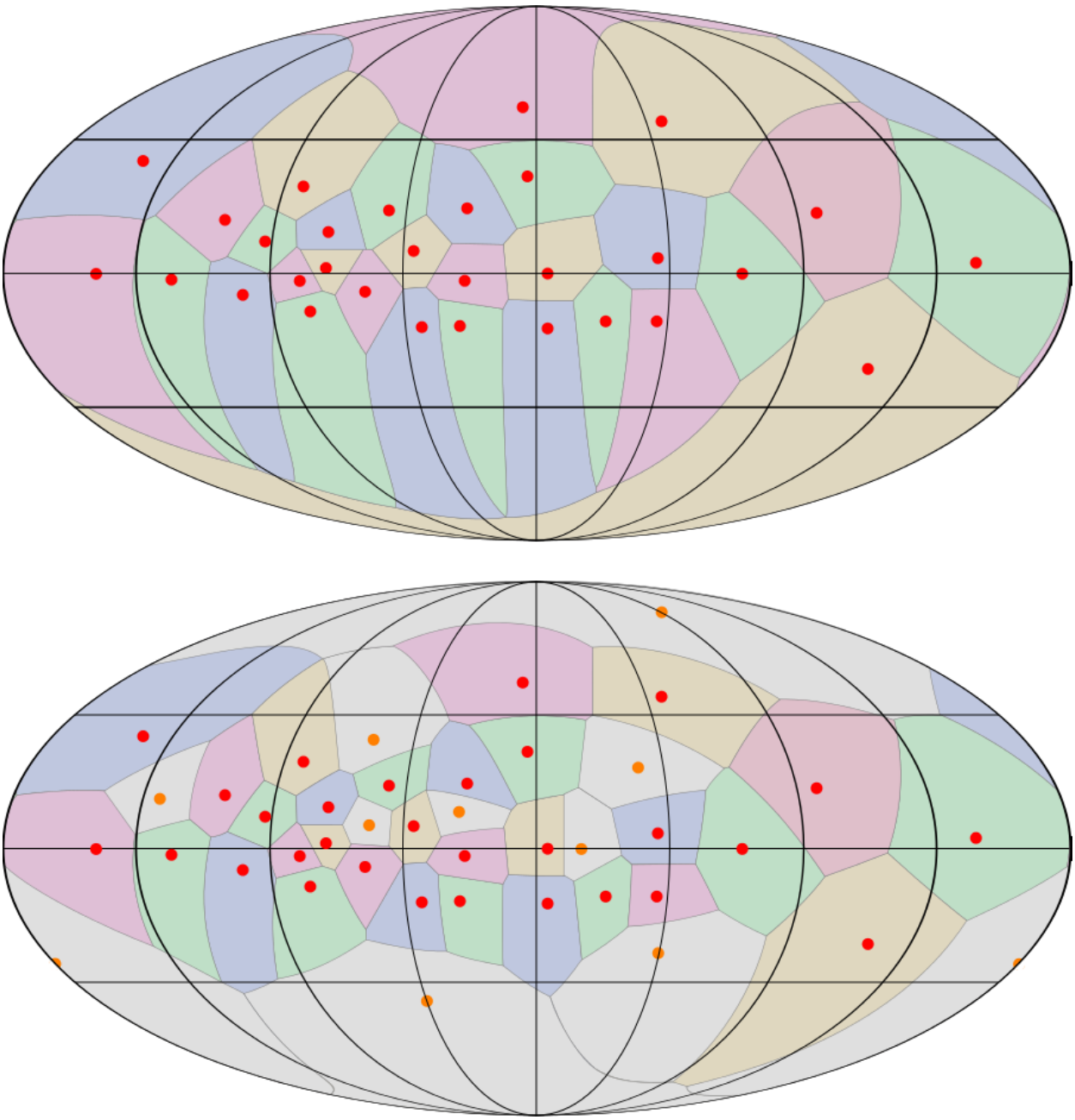


**Figure 64: Room Pointing targets during trials (top) and trials$^{+}$ (bottom); targets added for trials$^{+}$ are yellow dots on grey background; coordinate system lines at 45° increments**

The overall pointing error during, i.e., the angular distance between pointing target and produced pointing gesture, was $\bar{x} = 7.8°$, $\sigma = 4.9°$ for targets in front of the user (ventral targets, see Figure 65). The average angular distance between two targets was 22.8°, which means that in average pointing errors greater than 11.4° resulted in a wrong selection. While this value is not a

precise measurement for participant's pointing capabilities, it gives a general estimate about their pointing performance (see Figure 66, top).

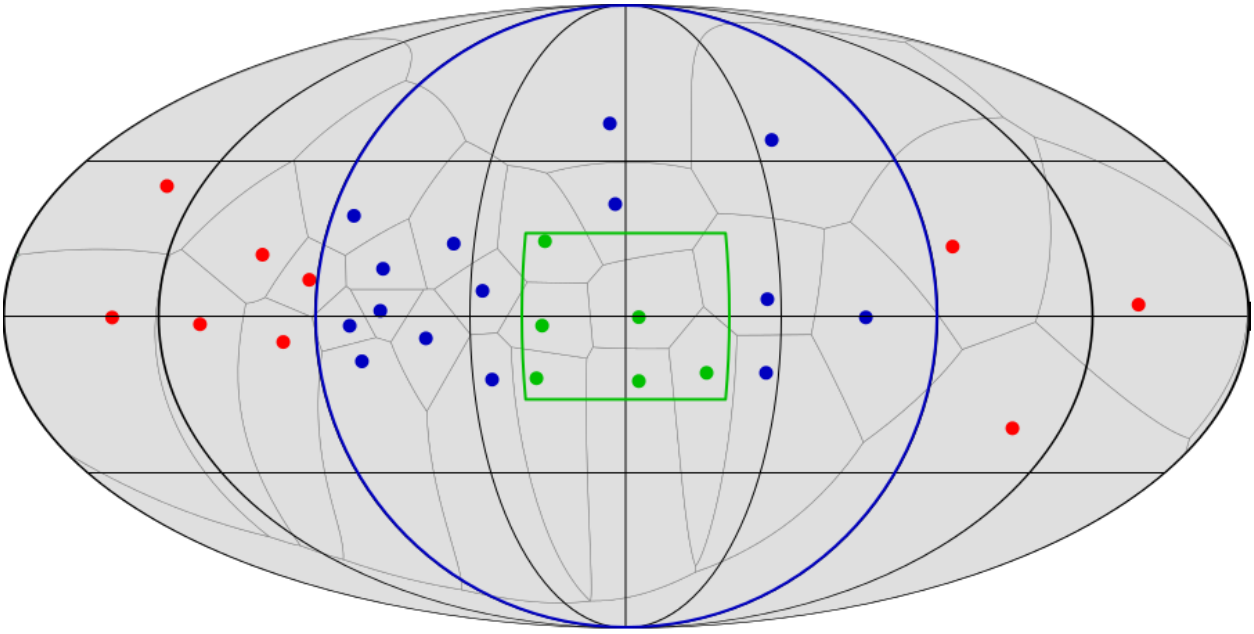


**Figure 65: Targets in front (blue) and within the central 60° × 60° frame (green)**

Figure 66 (bottom) shows just real-world proxies within the central 60° × 60° frame (pointing accuracy: $\bar{x} = 7.5°$, $\sigma = 4.3°$; selection accuracy: 89.2 %; average angular target distance: 24.2°). It is of the same size as the frame used in the evaluation of *Ray-casting Air-pointing* (Cockburn et al., 2011), though in my study there were 6 targets instead of 4. In order to assess the influence of target direction relative to the participant's viewing direction, I analyzed the difference in pointing accuracy between ventral targets ($30° < \varphi, \theta \leq 90°$, see Figure 65, blue) and targets within the central frame ($\varphi, \theta \leq 30°$, Figure 65, green). A paired-samples t-test did not reveal a significant difference ($t(109) = -0.06$, $p > .1$). This might suggest that people's pointing performance for the purposes of *Room Pointing* remains stable across their entire frontal hemisphere.

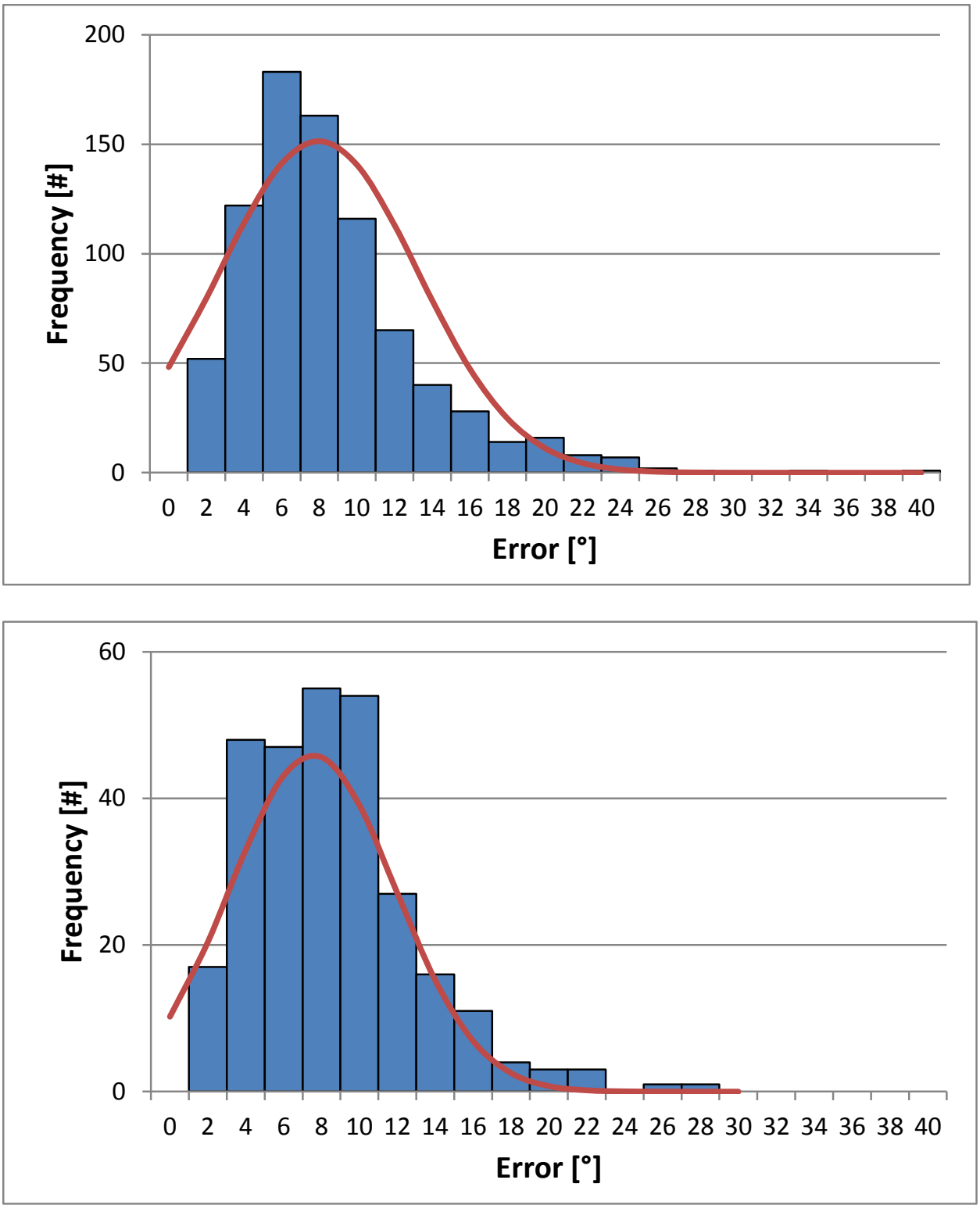



**Figure 66: Pointing errors for targets in front of the participant (top) and for all targets within ±30° of the user's view direction (bottom)**

### 5.5.4 Limitations of this Study

The goal of this study setup was creating an ecologically valid domestic environment for UbiComp interaction. I controlled for some variance in my experiment but ignored others, such as the amplitude of arm motion in *Screen Pointing* and *Room Pointing*. As mentioned before, I do not believe that my approach affects the validity of my work as I was less concerned about which technique is faster and more about whether they allow for similar performance.

#### *Use of Interaction Devices*

I called both *Screen Pointing* and *Room Pointing* device-free interaction, i.e. people can use them without holding an interaction device. In the user study, however, I let participants hold an interaction device for tracking their motion and confirming their selections. For both purposes, alternative techniques exist that keep people's hands free and allow confirmation without using physical buttons.

Tracking people's movements can be achieved without requiring any markers by systems such as the Microsoft Kinect. Out-of-the-box this system is, however, not as accurate as marker-based tracking, for example, through an optical tracker (see 4.1.2). While it is possible to achieve sub-millimeter tracking accuracy with multiple Kinect sensors (Ren, Liu, and Lim, 2013), the setup of such a system requires significantly more considerations than with an optical tracker and is therefore less feasible for the purpose of my studies.

A common device-free method for selection confirmation is "dwelling". With dwelling, people hold their arms still over the pointing target for a short dwell-time ($0.5\ s$) to confirm the selection (Wilson and Oliver, 2003). Another method that does not require any additional time is muscle-tracking. One can imagine to use the hand itself as a button and perform a quick thumb-tap on, for example, the curled-up middle finger during the pointing gesture. These muscle activities can be tracked by an electromyography-bracelets (EMG-bracelet) close to the elbow (Saponas, Tan, Morris, and Balakrishnan,2008).

In summary, the technology exists for tracking people's limb motion without the need for any markers and for making selections without the need to press a button or hold a device. These technologies, however, are experimental and oftentimes require a special setup to reach the accuracy and sensitivity of traditional IT-tracking systems or physical buttons. I decided to only

use established and well-understood technologies in my studies because the flaws of experimental technologies could have severely confounded my results: instead of measuring people's capabilities, I would have measured hardware limitations.

### *Comparability of Conditions*

There is one minor difference in the implementation of *Screen Pointing* and *Touch Scroll* and *Touch Flat*: in the *Screen Pointing*, the first button press opened the selection menu directly (see Figure 60, bottom), whereas with *Touch Scroll* and *Touch Flat*, the first touch opened the main menu (see Figure 57 left), from which participants had to open the selection menu (see Figure 57, center and right) with an additional tap. This means that *Touch Scroll* and *Touch Flat* required one more user action than *Screen Flat*. Indisputably, this additional action added to participant's completion time for both techniques. I argue, however, that this additional action does not change the overall results of my study for two reasons.

First, interacting with a smart phone would very likely require one more action than interacting with a large display, such as a TV screen. To reach the menu screen of a real HEI app running on a smart phone or tablet, people would have to grab the device, unlock it, navigate to the correct screen, and finally start the app. For the experiment, I assumed that people were already holding their device and that the HEI app can be started from the home screen. This still leaves a minimum of two actions: unlocking the phone and starting the app. To reach the menu screen of a real HEI app running on a large display, people have to orient themselves toward the screen and start the app, for example, by executing a certain gesture. Similar to before, I assumed that people were already oriented toward the display at the beginning of their interaction. This leaves only one required action: starting the app. Overall, this means that interaction through a smart phone or tablet generally requires one more action than interaction through a large display. In my experiment, I should have labelled this additional step "Unlock" instead of "Room Control" but it would not have changed the results.

Second, one might argue that it the unlocking-action is not necessary. In this case, selection times would decrease for both *Touch Scroll* and *Touch Flat*. I argue that the amount, however, would not affect my comparison between the four techniques. The additional tap on the "Room Control" button in *Touch Scroll* and *Touch Flat* very likely added only a fraction of a second to the overall completion time. This would mean that *Touch Scroll* would still be significantly

slower that the other three techniques, given people's slow overall speed with *Touch Scroll*. The difference between *Touch Flat* and *Screen Pointing*, in contrast, might have become non-significant. This, however, would not have changed my conclusion from this experiment since I never expected *Screen Pointing* to be faster than *Touch Flat*. I only hypothesized that people can use pointing-based interaction with the same levels of accuracy and speed than touch-based interaction.

### *Ceiling Effects*

The overall high selection accuracy for *Touch*, *Touch Flat*, and *Screen Pointing* could indicate the presence of a ceiling effect (see 5.4.2). This means that my study could not reveal significant differences in people's selection accuracy (independent variable) between selection techniques (dependent variable) because the selection task was not difficult enough for people to exhibit these differences. There are two arguments why this effect would not affect the results of this study. First, I would argue that there was no ceiling effect. Selection accuracy and speed are closely related through the concept of fluency (see 2.5.4), and it is likely that the difference in people's performance expressed themselves not in selection accuracy but in selection speed, a typical speed–accuracy trade-off (see 2.5.4). In this case, increasing the difficulty in the selection task, for example, by further reducing the size of icons would not have affected selection accuracy but selection speed instead. Second, I would argue that a ceiling effect would not affect the interpretation of the results and the validity of my hypotheses. My discussion is solely based on actually discovered differences. A ceiling effect, though being able to obfuscate significant differences, could not have produced invalid ones. Subsequently, neither presence nor absence of a ceiling effect would have reduced the findings of this study. It is, however, possible that a ceiling effect could have obscured significant differences between selection techniques other than the ones I found in my study. While these additional finding could have been interesting, they are not necessary in the context of the hypotheses for this study.

## 5.6 Conclusion

Interactions with smart environment can be vastly different, so that no single interaction technique will be able to excel in every possible scenario. Traditional navigation- and touch-based interfaces are lacking some key characteristics that can be desirable for interaction with smart environments, for example the possibility for device- and system-feedback-free

interaction. In this chapter, I confirmed that with pointing-based interaction techniques, people can still achieve performance levels comparable with traditional touch-based interaction techniques in terms of selection speed and accuracy. My results suggest that people are able to make fast and accurate selections using non-traditional, i.e. pointing-based, techniques. Compared to the traditional touch-based interaction techniques, pointing-based techniques offer additional advantages, such as device- and eyes-free interaction. This finding is important as it increases interaction designers' repertoire of potential techniques for HEI and gives them more possibilities to tailor user interfaces to the diverse range of use cases in smart environments.

In this chapter I showed that the benefit of device-free interaction is not diminished by reduced selection time or accuracy. In contrast, the advantage of room-based interaction, that is providing eyes-free interaction, is countered by significantly reduced selection accuracy. The use of mid-air full-arm pointing gestures in room-based interaction can help people with HEI in situations where they want to perform device-, system-feedback-, or, potentially, eyes-free interaction.

# Chapter 6 The Effect of Proxy Type on Memorability of Pointing-based Interactions

In Chapter 5, I showed that people can make selections with both full-arm pointing-based interaction techniques equally well as with traditional touch-based techniques while providing the extra benefit of device-free interaction (5.5.1). I also showed that the advantage of room-based interaction (eyes-free interaction) comes with reduced selection accuracy (5.5.1). Two possible reasons I hinted at in the previous chapter were the lack of system feedback about the currently selected proxy and people's potential problems with remembering where to point at, i.e. the association between digital command and real-world proxy-object: room-based interaction is memory-based, and people inherently have to rely on intrinsic visual and proprioceptive feedback alone (see 3.2.2). The question is whether it is possible to provide eyes-free interaction while sacrificing accuracy to a lesser degree than *Room Pointing*. An answer is closely related to another possible reason why *Room Pointing* suffered from lower selection accuracy: there might be a problem with the selection proxy type and design, and people have difficulties remembering the associations between digital artifact and real-world proxy item. Other interaction techniques, most notably the *Air Pointing* techniques (Cockburn et al., 2011), have successfully employed body-relative instead of real-world proxy objects. In this chapter, I set out to determine the influence of the selection proxy type on user performance and whether body-relative selection proxies are better suitable for feedback-free interaction with smart environments than real-world proxies. The main question I will answer is

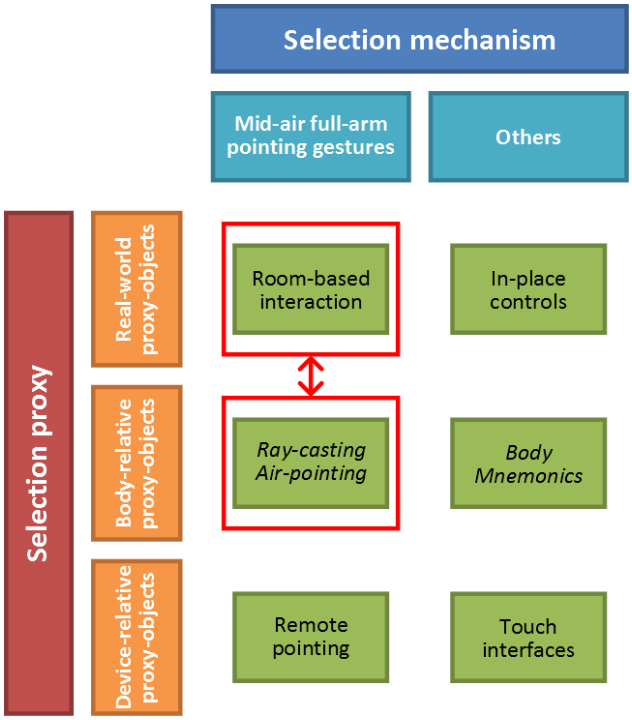



**Figure 67: Design space of study 2**

1. Would a different selection proxy type for mid-air full-arm pointing gestures increase people's performance?

## 6.1 Selection Proxy Types

Selection techniques in which people use their fingers or arms to select system functionality through natural pointing, have long been researched (see 2.2.3) and are now used commercially as well (e.g., Nintendo Wii, Microsoft Kinect). With *Virtual Shelves* (Li et al., 2009) and *Ray-casting Air-pointing* (Cockburn et al., 2011), researchers provided in-depth analyses of human performance for full-arm natural pointing selection techniques. In both studies, however, the authors use virtual, invisible regions in the environment as selection proxies (see Figure 70 and Figure 71 for an illustration). A possible verbal descriptor of a mapping between digital artifact and proxy region would be "turn light on" and "point toward 10° up and 30° right". These regions are located relative to the users' body, hence the term "body-relative". This means that whenever users change their location, the selection proxies move with them.

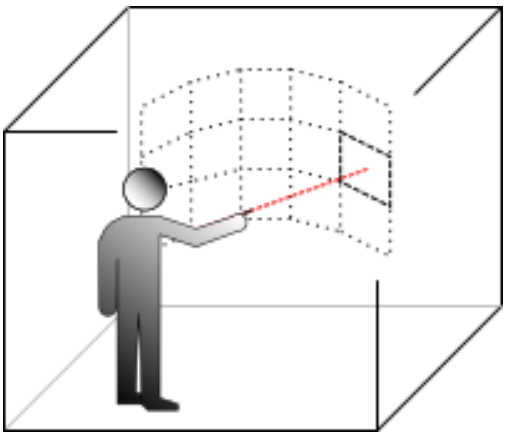


**Figure 68: *Ray-casting Air-pointing*: pointing-based interface using virtual body-relative proxy regions**

This characteristic sets these established selection techniques apart from room-based interaction. In *Room Pointing*, for example, selection proxies are relative to the environment itself and do not change no matter the location of the user.

Besides this key difference, both *Ray-casting Air-pointing* and *Room Pointing* show several similar characteristics: both selection techniques use mid-air full-arm pointing gestures for their selection mechanisms, both enable device- and system-feedback-free interaction, and both allow for eyes-free selections. On a cognitive level, however, both techniques are indeed different as

laid out in sections 3.2.2 and 3.2.3. This means that, although both techniques have similar properties, they work differently on the underlying level, and, therefore, might show differences in user performance.

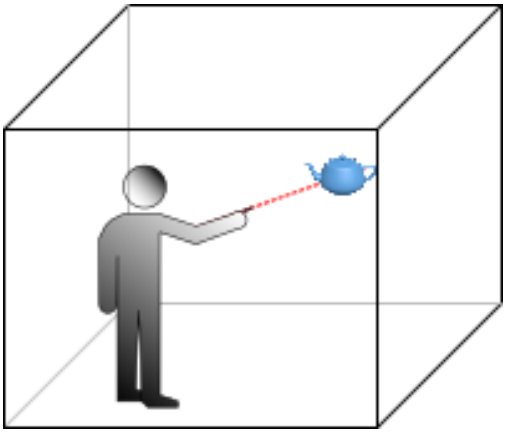


**Figure 69: *Room Pointing*: pointing-based interface using real-world proxy objects**

To show the influence of selection proxy type on user performance, I compare *Room Pointing* to a selection technique that uses body-relative proxy objects, such as *Virtual Shelves* and *Ray-casting Air-pointing*. I looked at the following five issues in particular:

- Learnability of proxy types: does the proxy type have an influence on how quickly and accurately people learn associations between digital artifacts and proxy objects
- Performance of proxy types: does the proxy type have an influence on how quickly and accurately people make selections
- Preference of proxy types: does the proxy type have an influence on people's subjective preference
- Usefulness of proxy type: does the proxy type have an influence on the usefulness of a selection technique in an everyday scenario

From the previous analyses of both *Room Pointing* and *Ray-casting Air-Pointing* (*RCAP*), I formulated four hypotheses:

1. Users can learn *Room Pointing* faster than *RCAP*
2. Users can initially make selections faster with *Room Pointing* than with *RCAP*
3. Users can initially make selections more accurately with *Room Pointing* than with *RCAP*
4. Given the advantages of *Room Pointing*, users will prefer *Room Pointing* over *RCAP*

## 6.2 Study Conditions

The following sections provide details on the implementation of the two study techniques.

### 6.2.1 Room Pointing

*Room Pointing* uses landmarks to store digital items, and users perform ray-casting-style pointing without system feedback to select items. The implementation of *Room Pointing* is the same as described in the previous study (see 5.2.3).

### 6.2.2 Ray-casting Air-pointing

**Figure 70: *Ray-casting Air-pointing* (Cockburn et al., 2011); virtual shelves are superimposed to illustrate the idea**

All *Air-pointing* techniques implemented by Cockburn et al. (Cockburn et al., 2011) use the metaphor of pigeonholes that users can either "reach" into or point at to select the digital item that is associated with the pigeonhole. Cockburn's study found that users performed best with a

simple 2D arrangement in combination with full-arm pointing gestures ("*Ray-casting Air-pointing*" or *RCAP*).

*RCAP* is similar to *Virtual Shelves* (Li et al., 2009), with the exception of the layout of the pigeonholes or shelves: the original *RCAP* arranged pigeonholes in a straight line in front of the user (Figure 70), whereas the shelves in *Virtual Shelves* formed a semi-circle around the user (Figure 71).

**Figure 71: My implementation of *RCAP*; inspired by *Virtual Shelves* (Li et al, 2009); virtual shelves are superimposed to illustrate the idea**

The difference between the original *RCAP* and *Virtual Shelves* is that the angular size of the pigeonholes in the original *RCAP* becomes smaller for the outer pigeonholes, whereas the angular sizes for the shelves remains constant in *Virtual Shelves* (Figure 72). This means that in *RCAP*, people's selection accuracy depends on the location of the selection proxy since pointing at smaller targets is more error-prone than pointing at larger ones (see 2.4.4).

To reduce the effect of the proxy location, I decided to use a semi-circular shelf-arrangement similar to *Virtual Shelves* in my study.

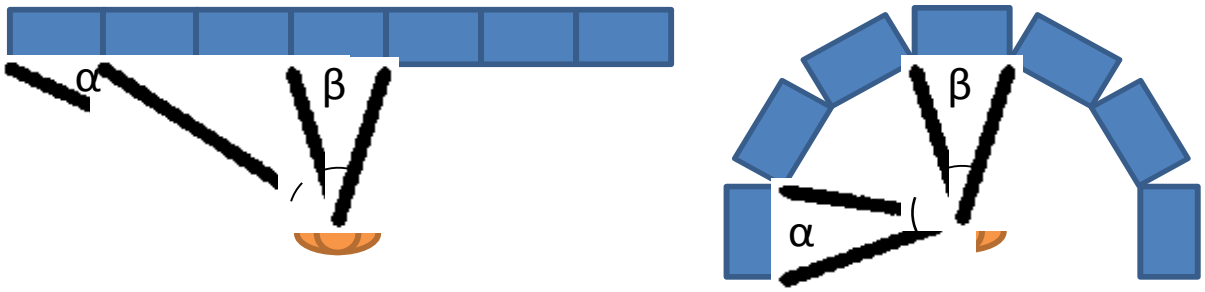



**Figure 72: Angular size of pointing targets in *RCAP* (left) and *Virtual Shelves* (right)**

The algorithm I used for my *RCAP* implementation is simple: shelves are defined by the four-tuple $s := \{\Psi_{min}, \Psi_{max}, \Theta_{min}, \Theta_{max}\}$. People then select a certain shelf $s$ if the input angles $in = \{\Psi_{in}, \Theta_{in}\}$ satisfy $\Psi_{min} < \Psi_{in} < \Psi_{max} \wedge \Theta_{min} < \Theta_{in} < \Theta_{max}$.

### 6.2.3 Moving through the Environment

The task-supporting nature of Human-Environment Interaction as described in 3.3.2, suggests that people might have to perform HEI in different locations throughout the environment depending on where the primary tasks takes place. This implies that people move around in the environment: they change their locations and their orientation. It appears that performing HEI under these variable circumstances would be common in a realistic scenario. Therefore, I decided to incorporate this effect in my user study. In the last phase of the experiment, I instructed participants to turn 90° to the right.

## 6.3 Experimental Setup

### 6.3.1 Study Design, Participants, and Apparatus

The study used a single-factor within-participant design with interaction technique as a two-level factor (*Room Pointing* and *RCAP*). The order of appearance was balanced using a Latin square.

I recruited 12 participants (8 male,6 female; ages $\bar{x} = 25.5$ years, all right-handed) from a local university. Participants received a $10 honorarium for participating in this one-hour-long study.

To perform the comparative study of *Room Pointing* and *Ray-Casting Air-Pointing*, I implemented a testing system that allowed users to use both techniques using free and unrestricted pointing movements. The system used eight NaturalPoint OptiTrack S250e IR-

tracking cameras to capture participants' location and pointing gestures. The IR-tracking system provided location and orientation information to my study system (see 4.2.7). The system was written in C# and ran on a standard Windows computer. To track participants' pointing gesture, I taped a small IR-reflector to their index and middle finger (for illustration see Figure 59 (right) in 5.3.3).

I set up both *Room Pointing* and *RCAP* in one section of the HCI research lab. Participants faced a $42''$ ($105\ cm$) TV screen that displayed the user interface. During all phases except the rotated condition (Trials$^{rot}$ 1 and 2), participants stood approximately $8.5'$ ($2.6\ m$) away from the screen. For Trials$^{rot}$1 and 2, I displayed the user interface on a $20''$ ($50\ cm$) computer screen that was approximately $4.5'$ ($1.4\ m$) away. This means that participants had a comparable viewing angle on the user interface throughout all phases, which include Trials 1 – 5 and Trials$^{rot}$ 1 and 2.

The study system logged a continuous stream of participants' locations and orientations as well as completion time for each trial, whether or not it was successful, and specific orientation details of the tracker when selections were made.

### 6.3.2 Digital Artifacts and Proxy Objects / Zones

I created two sets of 14 digital artifacts. One set was used for *Room Pointing* and the other one for *RCAP* in order to avoid learning-effects between conditions. The use of the item sets was counterbalanced between selection techniques. Although no item appeared on both sets, I tried to keep the sets comparable in order to reduce the effect of having two distinct sets.

I then created the virtual shelves for *RCAP*. I arranged them in a $7 \times 2$ semi-circular pattern (see Figure 71 for a realistic and Figure 73 for a conceptual representation). Subsequently, the shelves had $\Delta\Psi = 30°$ width. I decided to limit the height of each shelf to $\Delta\Theta = 60°$, so that the total size of each of the seven shelves was $\Delta\Theta = 120°$. I did so because for angles close to $\Theta = \pm 90°$, jitter of the human arm, wrist, and fingers makes it difficult to perform accurate and stable pointing gestures. Since there is no obvious mapping between digital items and virtual shelves, I tried to pair similar items together and stored them in the same rack, for example, "Simpsons" and "Game of Thrones" (both TV shows) and "volume up" and "volume down" (both device commands). For *Room Pointing*, I picked 14 landmarks to which I mapped the two sets of digital items. I pick these objects so that they would also form a $7 \times 2$ pattern similar to the one in

*RCAP*. The reason for this was that I wanted participants to perform similar pointing gestures for both selection techniques. I then mapped the digital objects from both lists to the 14 landmarks. (See Figure 74 for the final landmark setup.) Figure 73 and Figure 74 also illustrate the difference in shape between proxy object in *Room Pointing* (circular) and *RCAP* (rectangular). For *Room Pointing* I set the radius of landmarks to 34° to achieve the same combined shelf- and landmark-size as in *RCAP*.

**Table 6: Stimuli (digital artifacts)**

| Set #1 | Set #2 |
|---|---|
| Lights: dimmer | Volume up |
| Lights: brighter | Volume down |
| Game of Thrones | Breaking Bad |
| CBC Radio One | 97.3 Radio |
| Vacation photos | Family photos |
| mute / un-mute sound | TV on / off |
| New York Times | TIME Magazine |
| Facebook | Gmail |
| Call Jane | Call Steve |
| World of Warcraft | Grand Theft Auto IV |
| ShoutCast Hip-Hop | ShoutCast R'n'B |
| Netflix | Spotify |
| Rotten Tomatoes Top 10 | Roger Ebert's Movie of the Month |
| Simpsons | Family Guy |

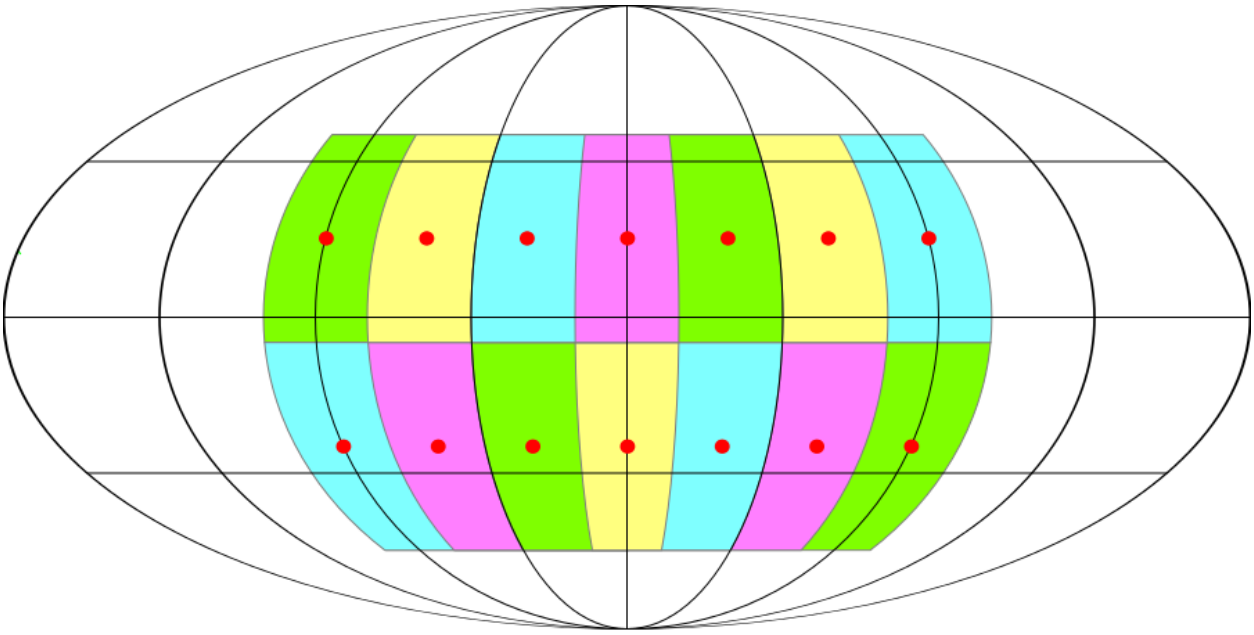


Figure 73: Ray-Casting Air-Pointing shelves; coordinate system lines at 45° increments

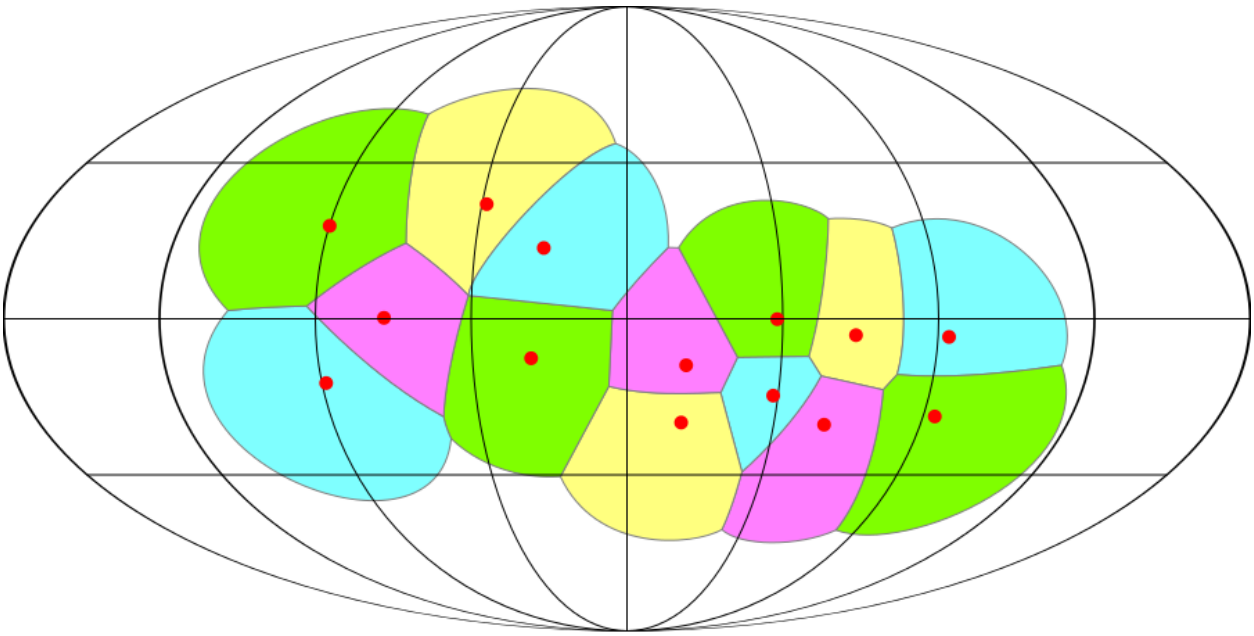


Figure 74: Room Pointing landmarks; coordinate system lines at 45° increments

Finally, I measured the mean pitch of all landmarks in *Room Pointing* ($\mu_\theta = -7.3°$) from the position in which the participants were supposed to stand and shifted the virtual shelves in *RCAP* downwards so that the targets in both selection techniques had the same mean pitch. This was to make the total distance travelled by the participant's arm comparable in both techniques.

### 6.3.3 Rotating Participants in the Environment

In a realistic smart domestic environment, people would move around and perform HEI from different locations. In this study, wanted to simulate this behavior. I therefore rotated participants by 90° clockwise in the final two blocks of each condition (trails$^{rot}$).

### 6.3.4 Study Conditions and Procedure

I asked participants to perform multiple selections of all 14 digital artifacts. A set of 14 selections was called a block; within a block, all 14 digital items were selected exactly once; the order within a block was separately randomized for all blocks. Overall, participants had to complete 15 blocks (demonstration, training, Trials 1, 2x training, Trials 2, 2x training, Trials 3, 2x training, Trials 4, Trials 5, Trials$^{rot}$ 1, Trials$^{rot}$ 2) for a total of 210 selections. For my evaluation, I only considered data collected during trial-phases (Trials and Trials$^{rot}$).

After the experiment, participants filled out one questionnaire asking for basic demographic data (e.g., age, gender, experience with full-arm pointing techniques) and one NASA-TLX form.

### 6.3.5 Data Analyses

To determine the effect of the selection technique on participant's performance, I analyzed the trial phase and the trial$^{rot}$ phase separately. For the main trial phase, the analysis consisted of a $2 \times 5$ (technique by block) RM-ANOVA; for the rotated phase $2 \times 2$ (technique by block). Post-hoc tests used Bonferroni correction for all between-block and between-technique analyses. I evaluated the questionnaire data using Wilcoxon Signed Rank tests.

## 6.4 Results

### 6.4.1 Accuracy

#### *Main Testing Phase (Trials 1 – Trials 5)*

For the main testing phase, there was a main effect of technique on accuracy ($F(1,11) = 19.6, p < .001$), with participants having significantly higher accuracy with *Room Pointing*

($\bar{x} = 92.3\ \%, SE = 2.0\ \%$) than with *RCAP* ($\bar{x} = 72.1\ \%, SE = 5.0\ \%$). There also was a main effect of block on accuracy with participants ($F(4,44) = 13.8, p < .001$).

**Table 7: Mean selection accuracy and standard error**

| | ***Room Pointing*** | ***Ray-casting Air-pointing*** |
|---|---|---|
| **Trials 1** | $87\ \% \pm 4\ \%$ | $51\ \% \pm 7\ \%$ |
| **Trials 2** | $94\ \% \pm 3\ \%$ | $69\ \% \pm 6\ \%$ |
| **Trials 3** | $91\ \% \pm 2\ \%$ | $76\ \% \pm 6\ \%$ |
| **Trials 4** | $96\ \% \pm 2\ \%$ | $80\ \% \pm 6\ \%$ |
| **Trials 5** | $94\ \% \pm 3\ \%$ | $84\ \% \pm 7\ \%$ |
| **Trials$^{rot}$ 1** | $89\ \% \pm 6\ \%$ | $58\ \% \pm 6\ \%$ |
| **Trials$^{rot}$ 2** | $89\ \% \pm 5\ \%$ | $60\ \% \pm 6\ \%$ |

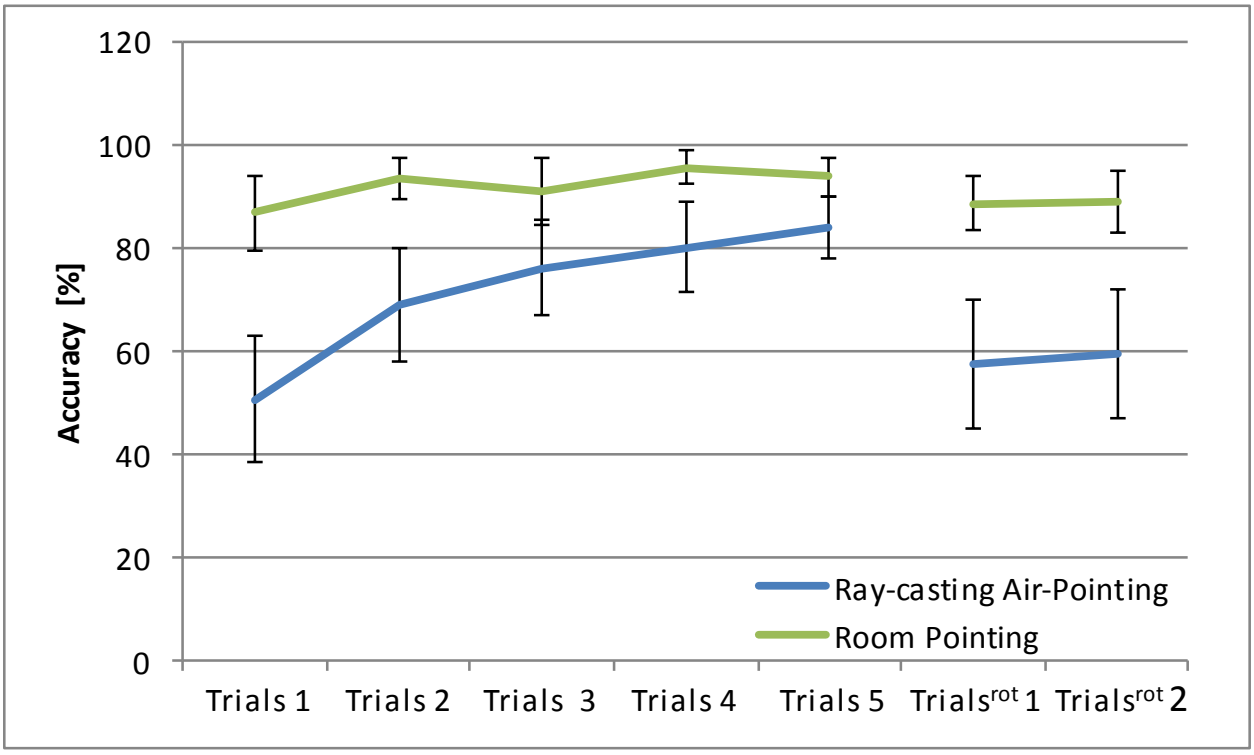



**Figure 75: Overall accuracy**

Finally, there was a significant interaction between selection technique and block number ($F(2.5,27.7) = 4.5, p < .05$). Pair-wise block-analysis revealed that participants were significantly more accurate with *Room Pointing* from Trials 1 through Trials 4 ($p < .001$, $p < .01$, $p < .05$, $p < .05$). There was no significant difference for Trials 5 ($p > .2$) (see Figure 75). When comparing the first (Trials 1) and the last (Trials 5) block only for each selection technique, I found that performance improved with *RCAP* ($p < .01$), but did not with *Room Pointing* ($p > .05$). As shown in Figure 75, this effect may be due to the high initial selection accuracy with *Room Pointing*.

#### *Effect of Rotation (Trials 5 – Trials$^{rot}$ 1)*

There was a main effect of technique on accuracy ($F(1,11) = 11.1, p < .01$) with participants more accurate with *Room Pointing* than *RCAP*. There was also a main effect of block on accuracy ($F(1,11) = 15.2, p < .01$), with participants being more accurate before rotation than after rotation. Last, there was a significant interaction between technique and block ($F(1,11) = 22.7, p < .001$).

An analysis between blocks Trials 5 and Trials$^{rot}$ 1 revealed that participant experienced a significant drop in accuracy when using *RCAP* after rotation ($p < .01$). However, with *Room Pointing*, participants did not experience a drop in accuracy ($p > .05$) (see Figure 75).

#### *Recovery from Rotation (Trials$^{rot}$ 1 – Trials$^{rot}$ 2)*

When examining the trials after rotation there was a main effect of technique on selection accuracy ($F(1,11) = 36.4, p < .001$), with selection accuracy being higher for *Room Pointing* than for *RCAP*. There was no a main effect of block on selection time ($F(1,11) = 0.2, p > .5$) and no interaction between selection technique and block number ($F(1,11) = 0.1, p > .5$).

### 6.4.2 Completion Time

#### *Main Testing Phase (Trials 1 – Trials 5)*

During the main testing phase, there was no main effect of technique on completion time ($F(1,11) = 3.3, p > .05$). This means that overall, participants were not significantly faster with either *Room Pointing* ($\bar{x} = 2.2\ s, \sigma = .24\ s$) or *RCAP* ($\bar{x} = 2.9\ s, \sigma = .32\ s$). There was, however, a main effect of block on completion time ($F(1.3,14.0) = 22.2, p < .001$). Trials 1

was significantly slower than Trials 2 through 5 ($p < .01$), and Trials 2 significantly slower than Trials 4 and 5 ($p < .05$). There was no interaction between block and technique ($F(1.5,16.4) = 0.1, p > .5$).

**Table 8: Mean completion time and standard error**

| | ***Room Pointing*** | ***Ray-casting Air-pointing*** |
|---|---|---|
| **Trials 1** | $3.4\ s \pm 0.5\ s$ | $4.0\ s \pm 0.6\ s$ |
| **Trials 2** | $2.4\ s \pm 0.3\ s$ | $3.0\ s \pm 0.4\ s$ |
| **Trials 3** | $1.9\ s \pm 0.2\ s$ | $2.6\ s \pm 0.3\ s$ |
| **Trials 4** | $1.8\ s \pm 0.1\ s$ | $2.4\ s \pm 0.2\ s$ |
| **Trials 5** | $1.7\ s \pm 0.1\ s$ | $2.2\ s \pm 0.2\ s$ |
| **Trials$^{rot}$ 1** | $2.0\ s \pm 0.1\ s$ | $2.6\ s \pm 0.2\ s$ |
| **Trials$^{rot}$ 2** | $1.9\ s \pm 0.1\ s$ | $2.3\ s \pm 0.2\ s$ |

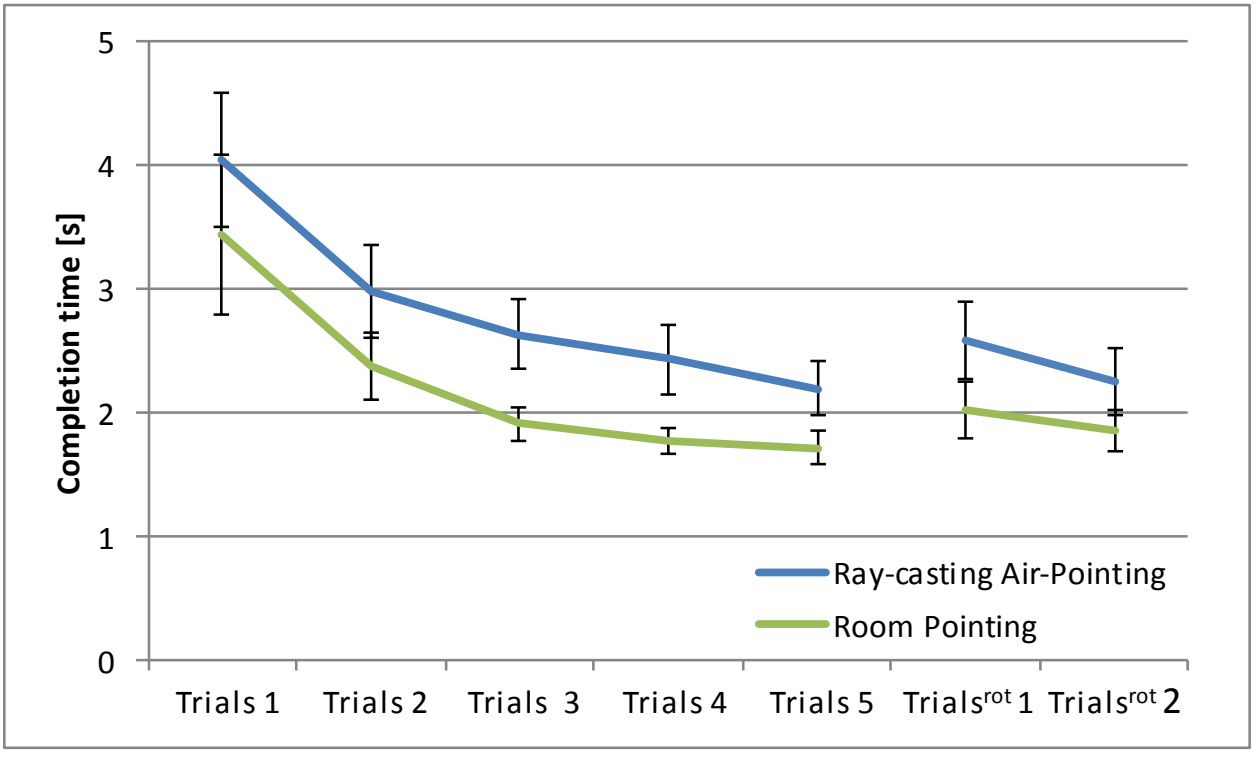



**Figure 76: Overall selection time**

#### *Effect of Rotation (Trials 5 – Trials$^{rot}$ 1)*

There was a main effect of technique on completion time ($F(1,11) = 10.8, p < .01$), with *Room Pointing* performing significantly faster than *RCAP* over both blocks. There was also a main effect of block ($F(1,11) = 16.0, p < .01$). For both techniques, average selection times increased after rotating participants (Trials$^{rot}$ 1) (*RCAP*: $+0.39\ s$; *Room Pointing*: $+\ 0.31s$). Figure 76 illustrates the slight increase for selection time during the rotated trial phase. There was no observed interaction effect between technique and block ($F(1,11) = 0.4, p > .05$).

#### *Recovery from Rotation (Trials$^{rot}$ 1 – Trials$^{rot}$ 2)*

When examining the trials after rotation there was a main effect of technique on completion time ($F(1,11) = 8.2, p < .05$), with completion times being lower for *Room Pointing* than for *RCAP*. There was also a main effect of block on completion time ($F(1,11) = 7.1, p > .05$), with participants completion times slightly lower in Trials$^{rot}$ 2 than in Trials$^{rot}$ 1. There was no interaction between selection technique and block number ($F(1,11) = 0.7, p > .05$).

### 6.4.3 Subjective Measures

I asked participants to rate their experiences with each technique along several dimensions (see Figure 77 and Figure 78). Overall, I found that participants rated *Room Pointing* as being easier to use ($p < .01$) and easier to learn ($p < .01$) than *RCAP*. Participants also rated *RCAP* as being more frustrating to use ($p < .05$) than *Room Pointing*.

When asked to directly compare *RCAP* with *Room Pointing*, participants strongly preferred *Room Pointing*. Out of twelve participants, 11 felt they were more accurate ($\chi^2(1,12) = 8.3, p < .01$) and 10 to be faster ($\chi^2(2,12) = 13.5, p < .001$) with *Room Pointing.* Ten participants of twelve found the mappings between digital items and landmarks easier to learn ($\chi^2(1,12) = 5.3, p < .05$) and easier to remember ($\chi^2(1,12) = 5.3, p < .05$) than between digital items and virtual shelves; 9 found RCAP easier to use ($\chi^2(2,12) = 9.5, p < .01$), and 10 overall preferred *Room Pointing* over RCAP ($\chi^2(1,12) = 5.3, p < .05$).

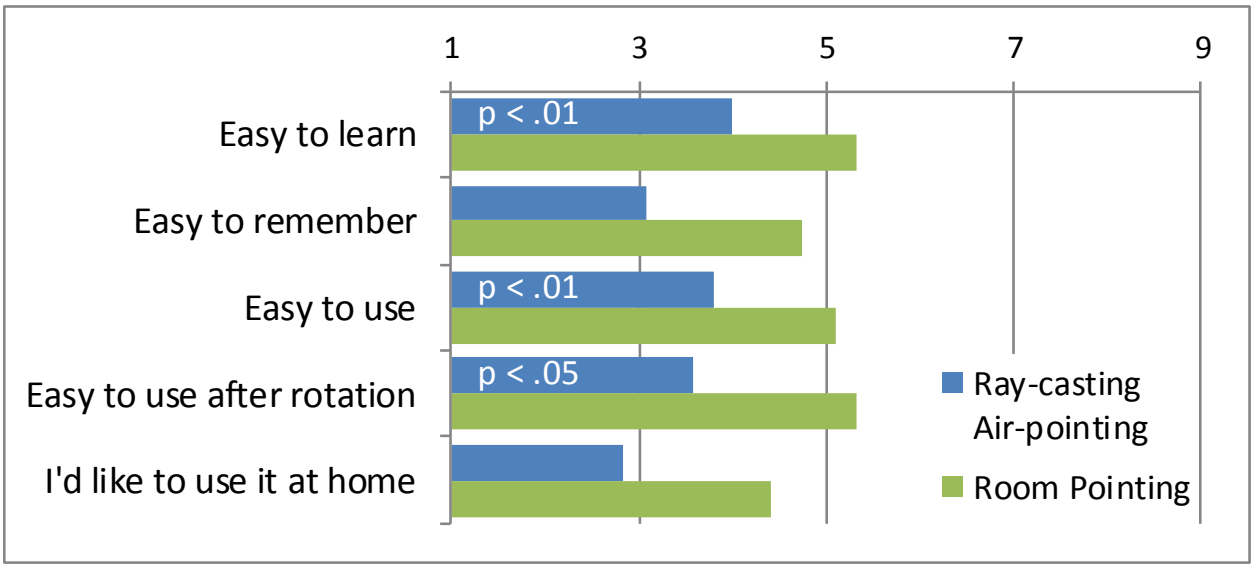



**Figure 77: Participant preference (higher is better); *p*-values given where difference significant**

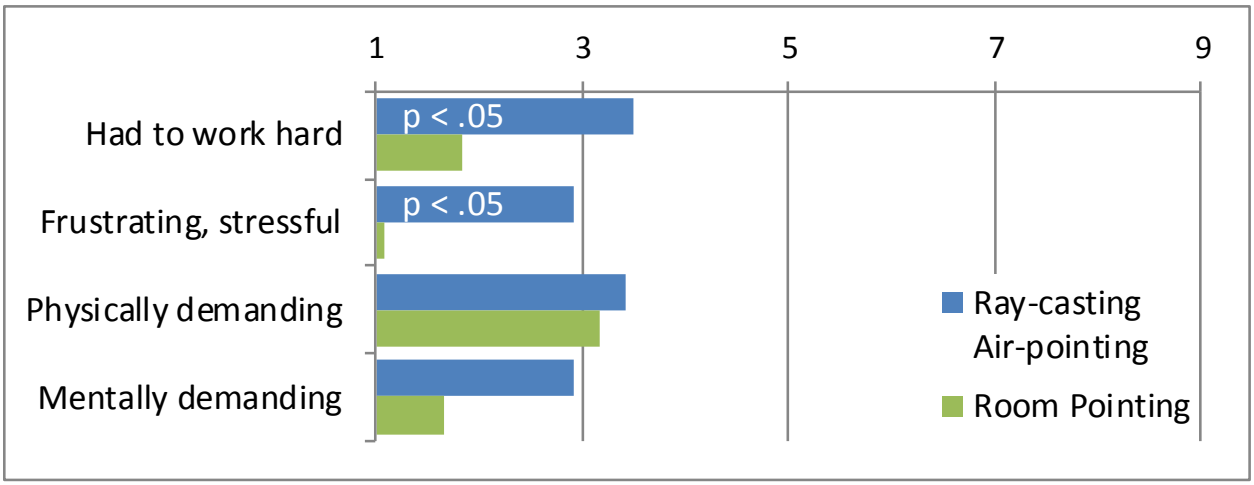



**Figure 78: Participant preference (lower is better); *p*-values given where difference significant**

## 6.5 Discussion

In the discussion, I first review the four hypotheses that I formulated at the beginning of this chapter. After this, I will discuss additional findings from the results of the experiment.

### 6.5.1 Review of the Main Hypotheses

From the previous analyses of both *Room Pointing* and *Ray-casting Air-Pointing* (*RCAP*), I formulated three hypotheses:

1. Users can learn *Room Pointing* faster than *RCAP*
2. Users can initially make selections faster with *Room Pointing* than with *RCAP*
3. Users can initially make selections more accurately with *Room Pointing* than with *RCAP*
4. Given the advantages of *Room Pointing*, users will prefer *Room Pointing* over *RCAP*

#### *Faster Learning through Semantic Memory*

After only two blocks, participants already achieved 87 % accuracy in *Room Pointing*. At this point, accuracy for *RCAP* barely exceeded 50 %. Furthermore, participants were already faster with *Room Pointing* compared to *RCAP* (*Room Pointing*: 3.4 $s$; *RCAP*:4.0 $s$). This confirms my first hypothesis. It also shows that people can learn *Room Pointing* with its underlying "digital item ↔ landmark"-associations better than *RCAP* with its underlying "digital item ↔ abstract pointing gesture"-associations. These results support existing research from psychology that people have less difficulties storing information in semantic memory than in procedural memory. In general the results of this study suggest that designers of selection techniques should carefully assess what memory systems their technique is going to rely on and consider using semantic memory if users should reach high proficiency after a short training only.

#### *Faster Selection with World Pointing*

This study demonstrates that people can perform selections with *Room Pointing* significantly faster than *RCAP*. At the end (Trials 5), participants were 0.5 $s$ faster with *Room Pointing* than with *RCAP* (*Room Pointing*: 1.7 $s$; RCAP: 2.2 $s$). Overall I showed that *Room Pointing* is significantly faster than *RCAP* even after 30 minutes of training with both techniques. This confirms my second hypothesis. Due to time limitations of my study (1 hour) and the fact that completion times for both techniques were still decreasing, I cannot predict the minimum selection time for fully trained participants. The data clearly supports, however, some assumptions about my cognitive *Room Pointing* analysis, namely that the first operator (recall of the association between system command and real-world proxy), does not significantly increase selection time (see A – C in 3.2.2). Having the additional operator for resolving the added level of indirection in room-based interaction, does not imply that it has to be slower than *RCAP*.

Again, I suggest that selection technique designers have to be considerate about the cognitive complexity of their technique and should not disregard any idea simply based on the number of operator but consider the complexity of each operator as well.

### *More Accurate Selections with World Pointing*

Only after 10 blocks of using *RCAP* (until Trials 5) did participants reach a similar level of accuracy as with *Room Pointing*, though the average accuracy for *RCAP* at the end of the main trial phase was still lower than the initial accuracy for *Room Pointing* (*Room Pointing*: 87 %; *RCAP*: 84 %). This confirms my third hypothesis. Due to time limitations of my study (1 hour) and the fact that selection accuracy for *RCAP* was still increasing, I cannot predict the maximum selection accuracy for fully trained participants. The data clearly supports, however, some assumptions about my cognitive *Room Pointing* analysis. First it shows that either the process of translating between relational and procedural information in *RCAP* lacks precision or that people have difficulties retrieving the correct verbal descriptor (response) for a given stimulus. Likewise, it shows that people have no such difficulties when recalling the association between system command (stimulus) and real-world proxy object (response) in *Room Pointing*.

As mentioned earlier, there are two possible reasons for making selection errors: pointing and recall errors. Unfortunately, it is difficult to distinguish between these two without active participant involvement (e.g., asking participants to say the proxy object out aloud before making a selection), and any active involvement inherently affects task completion time negatively. Instead, looking at the distribution of pointing errors can give insights about the reason why they occur.

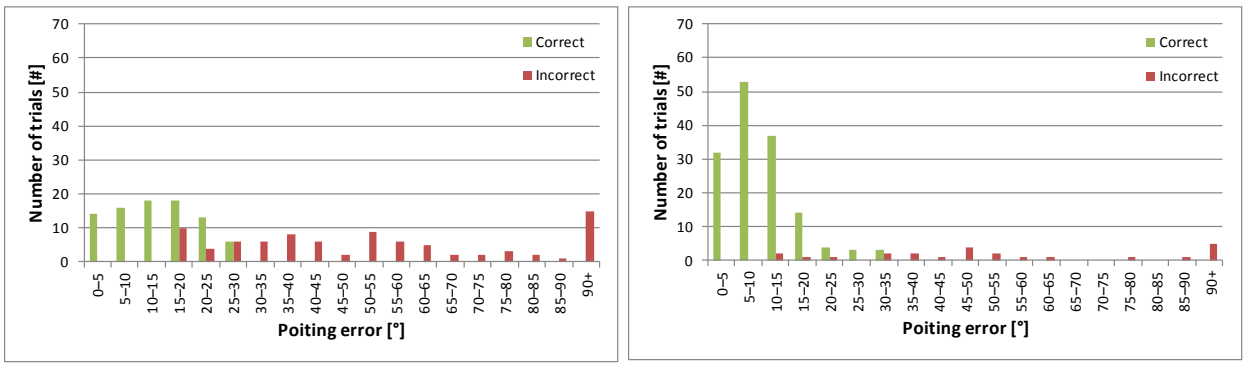



**Figure 79: Pointing errors during Trials 1 for *RCAP* (left) and *Room Pointing* (right)**

Figure 79 shows the absolute pointing errors during the first trial phase (Trials 1) for *RCAP* and *Room Pointing*. As a reminder, in RCAP targets are 30° across, in Room Pointing, they have a diameter of 34° or less. Given the general accuracy of human pointing gestures (see 2.4.4) and also the unfamiliarity of participants with full-arm pointing-based interaction, it is still safe to assume that incorrect selections with an error of, presumably, above 30° are almost entirely caused by recall errors and not pointing errors. For *RCAP*, 27 % of all performed selections fell into this category (23 % incorrect selections with an error below 30°, 50 % correct selections); in contrast, for *Room Pointing* only 9 % of the selections I would categorize as incorrect due to recall errors (5 % incorrect selections with an error below 30°, 86 % correct selections). These numbers strongly suggest that people have problems remembering the association between system command and proxy gesture (the first operator, see 3.2.3) as well as accurately recalling the details about the proxy gesture (the second operator, see 3.2.3). During the last regular trial phase (Trials 5), participants' performance generally increased. Now, only 5 % of the pointing errors in *RCAP* are most likely caused by incorrect recall (13 % incorrect selections with an error below 30°, 82 % correct selections). For *Room Pointing*, the numbers are $< 1$ % (incorrect recall), 5 % (incorrect selections with an error below 30°), and 95 % (correct selections) (see Figure 80).

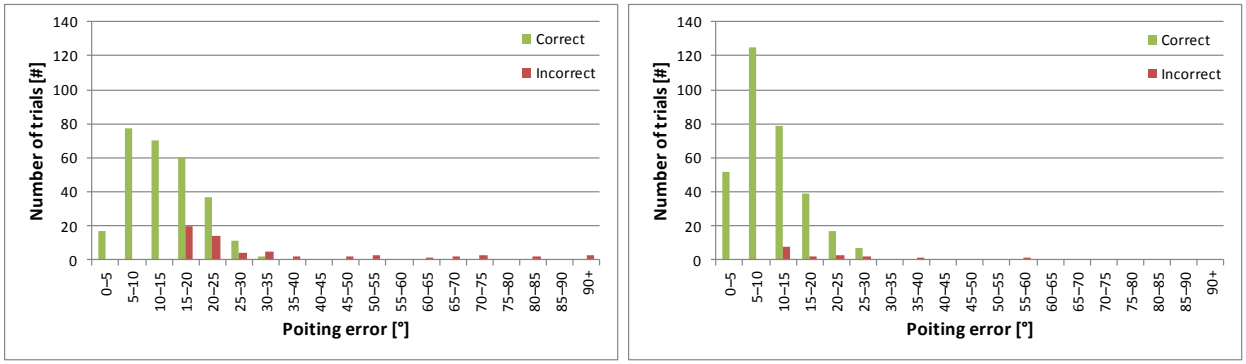



**Figure 80: Pointing errors during Trials 4 and 5 for *RCAP* (left) and *Room Pointing* (right)**

The data from Trial 5 also shows that after practice, participants overall produced lower pointing errors for all correct pointing gestures in *Room Pointing* ($\mu = 10.4°$) than in *RCAP* ($\mu = 13.6°$); an independent-samples T-test shows this difference to be significant ($t(560) = -6.2, p <$

0.01). Since the target areas for both *Room Pointing* and *RCAP* are of comparable size and, thus, do not necessitate people to be more accurate in one of the techniques, it is reasonable to assume that people are generally more accurate when pointing at a real-world object (*Room Pointing*) than at a virtual, invisible target zone (*RCAP*). Again, I implicitly predicted this behavior previously when arguing for a loss of pointing accuracy due to translating between relational and procedural memory (see 3.2.3).

#### *Users prefer World Pointing*

A significant number of participants found *Room Pointing* easier to learn and to use and less frustrating than *RCAP*. Overall, 10/12 participants preferred *Room Pointing* over *RCAP*, 2/12 *RCAP* over *Room Pointing*. Participant preference ratings sometimes appear to be biased toward novel interaction techniques. In this study, however, I believe that not to be an issue as both *Room Pointing* and *RCAP* were novel to all participants. Given the low accuracy of *RCAP* through the experiment, I are not surprised that participants were more frustrated with *RCAP* and subsequently preferred *Room Pointing*.

### 6.5.2 Effect of Rotating Participants

I did not have a clear hypothesis about the magnitude of the effects of rotating participants in terms of selection speed and accuracy. I assumed that selection time for *Room Pointing* should increase, while *RCAP* should remain unaffected. What I found instead was a major drop in accuracy for *RCAP* during Trial$^{rot}$. I cannot explain this finding conclusively but I can try giving a possible explanation.

I assume that participants did not conceptualize *RCAP* the way I expect them to. For me, the underlying metaphor of *RCAP* is virtual shelves or invisible pigeon holes. In the study, however, numerous participants conceptualized the virtual, invisible proxy zones through real-world objects that happened to be inside the zone. Essentially, participants turned *RCAP* into *Room Pointing*. Numerous quotes about the memorization strategy from the post-technique questionnaire bore witness of this behavior:

- "I associated positions of digital objects with certain objects in the room" (P1)

- "At first, I associated shelves with real-world objects; but eventually, I was able to fine-tune my accuracy and stopped paying attention to objects, just concentrated on the area I was pointing to" (P3)
- "I just tried to correlate objects seen around with the names of digital objects. For instance, for 'TV on or off' I memorized the remote control placed over the table" (P5)
- "I had to remember land marks in the room for each of the locations of the digital objects. When I was rotated 90°, I didn't get the help of landmarks I have created. It was just a guessing game. But I was sure in which shelf the object was lying." (P6)
- "Rather than remember shelves, I mentally assigned positions to the names. ex: Grand Theft Auto is played on the TV. Name the white box 'Steve' etc. After rotating I couldn't do this any more" (P7)
- "I used imagination to remember things. For example, I pointed a table for family guys because family are usually get together for dinner. I used a sofa for New York Times because people read papers in a sofa." (P10).

I should note here that P1, P3, P5, and P7 saw *RCAP* after *Room Pointing*, which could mean that the mental model of these participants was primed. Two of the participants, P6 and P10, however, saw *RCAP* first and still adopted a mental model based on real-world proxy objects. In contrast, none of the participants who saw *RCAP* first used invisible, virtual proxy zones in *Room Pointing*.

This behavior would give a good explanation on the drop in performance after I rotated my participants. After being rotated, the associations between virtual target zones and real-world objects become invalid, and since participants relied on the real-world object in their associative chain, they cannot perform their pointing gesture as accurately as before. The reason why participants did not conceptualize *RCAP* as expected remains unclear. It is possible that some participants might have been primed by the metaphor of "virtual shelves". I explicitly stated in the instructions that the virtual shelves rotate with the body and I contrasted *RCAP* and *Room Pointing* in the crucial difference that for former, participants still had to point at the same real-world proxy-object as before, while for the latter, participants had to perform the same physical gesture as before. However, some participants might have conceptualized the virtual shelves of *RCAP* to be static and therefore inert to rotation, like actual physical shelves.

Nonetheless, the results suggest that people have a tendency to naturally employ real-world objects as aids for skilled behavior. This is not surprising as existing theory predicts this behavior: the semantic nature of spatial information (see 2.5.3) supports beginners in memorizing procedural information during its initial learning phase (semantic phase, see 2.5.4). In squash , for example, players aim their three-wall boast shots at real-world landmarks outside of the court. Observing this behavior in this study showcases the advantage of using real-world objects as selection proxies and simultaneously hints at a potential danger to techniques like *RCAP*, which requires users to ignore and disregard the landmarks in the environment.

### 6.5.3 The influence of Proxy Types on Learnability and Selection Accuracy

The results from this experiment show that the type of proxy objects strongly influences people's learning curve and performance with an interaction technique. Since both *RCAP* and *Room Pointing* use the same mid-air full-arm pointing gestures (i.e., use the same final operator, compare 3.2.2 and 3.2.3), these performance differences must therefore be caused by the proxy object recall. *RCAP*'s comparably higher percentage of large errors (i.e., incorrect selections with errors above 30°) indicates that people have problems recalling associations between system commands and selection proxies (compare 3.2.2 and 3.2.3, first operators). Although this experiment does not provide a single reason for this problem, existing theory provides a plausible explanation: the lower amount of meaning between system command and proxy objects in *RCAP* compared to *Room Pointing* makes remembering more difficult. Similarly, *RCAP*'s comparably higher percentage of small errors (i.e., incorrect selections with errors below 30°) indicates that people have problems accurately recalling details about the proxy object (compare 3.2.2 and 3.2.3, second operators). Again, this experiment does not provide a single reason for this problem, though existing theory provides a plausible explanation: people cannot remember the different types of proxy details between *RCAP* (proprioceptive and visual impressions) and *Room Pointing* (spatial and visual impressions) equally accurately, which in turns limits people's capabilities for performing an accurate pointing gesture during the final operator. Overall, these results indicate that my analyses in sections 3.2.2 and 3.2.3. are plausible, that the proxy type has a significant influence on people's performance with and interaction technique, and that the real-world objects used in room-based interaction are of a more favorable proxy type than the body-centric virtual proxy regions used in the *Air Pointing* techniques.

### 6.5.4 Limitations of this Study

One limitation of this study was that it took place in a lab environment, which is considered to be ecologically valid than a real-world deployment. While the lab environment allowed for better control of external factors on performance data, it limits the significance of the results in a real-world scenario.

## 6.6 Conclusion

Human-Environment Interaction with mid-air full-arm pointing gestures as interaction mechanisms has many advantages, such as device- and system-feedback-free interaction. What type of selection proxy should be combined with these pointing gestures, however, was an unanswered question. In this chapter, I showed that real-world proxy objects can be a better alternative than the previously suggested virtual, invisible target zones. With this study, I confirmed that people can learn interaction techniques that primarily build upon semantic memory faster and easier than the ones that predominantly use procedural memory. Finally, I showed that with slightly more training, people can use *Room Pointing* and room-based interaction more accurately than my previous study suggested. The use of real-world proxy objects in room-based interaction can help people learn HEI faster and perform HEI more accurately.

## Chapter 7 Room Pointing as Tool for Creating Awareness

In the previous chapters, I showed that room-based interaction has advantages over navigation-based user interfaces and over pointing-based interaction techniques that use body-relative proxy objects (i.e. *Air Pointing* techniques). Both of my studies solely focused on the interaction between a single user and the digital system. Human-Environment Interaction, however, oftentimes include a social component as people share, for example, their smart domestic environments with others: their friends, partners, and relatives. There are situation in which these people can have an interest in knowing when someone else interacts with the smart environment. For example, bystanders might want to know if someone is about to turn on the lights, change the TV station, or do a certain move in a video game that is controlled through room-based interaction. The question that arises here is whether room-based interaction has the potential of making interactions with a smart environment better visible to co-located people than traditional navigation-based interaction. A reasonable assumptions is that gesture size, i.e. the amount of physical motion in a gesture, is one of the factors that determines the observability and identifiability of a gesture: larger gestures should be better observable that smaller ones. Figure 81 hints toward another important factor: the spatial orientation between actor and observer. In this chapter, I set out to determine the influence of gesture size and orientation between actor and observer of the identifiability and observability of gestures. The main questions I will answer are

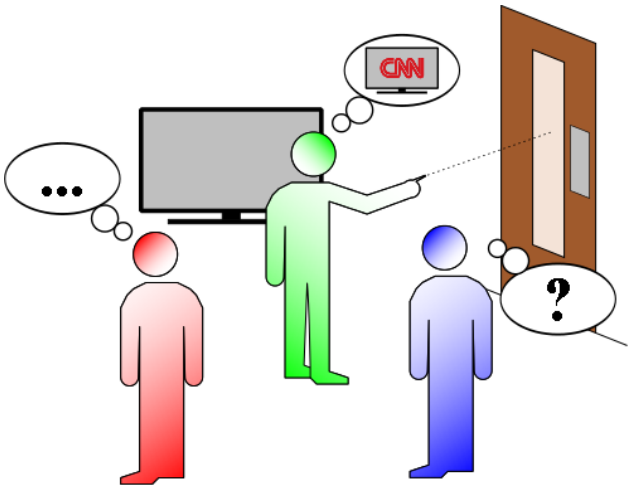



**Figure 81: Actor (green) and bystanders (red: not observing the interaction; blue: observing but not identifying the interaction)**

1. How large is the influence of gestures size on its observability and identifiability?
2. How does the mutual orientation between actor and observer influence gesture observability and identifiability?

## 7.1 A study of gesture observability

Gesture observability within the context of interacting with smart environments in domestic settings fits well within the theme of my dissertation. The overall topic, however, has a much broader application area and has been discussed in the design community for a long time (see 2.2.4). Subsequently, I want to take a broader approach in this chapter and not limit the applicability of my results to domestic environments but to all gesture-based human-computer interaction.

It is generally advantageous to be aware of the activities and interactions of others when working on the same digital system. Having group- or workspace-awareness normally improves the people's efficiency. This is true for co-located and remote, collaborative and competitive, and loosely and tightly coupled activities (see 2.2.4). A common problem for creating awareness is that people do not observe other's input or the changes that this input causes in the system (lack of feedthrough). As a result, many techniques amplify system input and system changes in order to draw people's attention and increase awareness (see 2.2.4). These techniques, however, often work only when people are observing the same part of the shared workspace, for example, the same interactive digital table. They do not provide a solution in situations where people are carrying out loosely coupled work in a co-located setting, for example, three people working collaboratively on an interactive table, a wall-mounted display, and a tablet.

A recent development, and one that could potentially improve the observability of system input, is the rise of gestural interaction techniques. Finger-based gestures are now common on hand-held touch-screen devices, such as smart phones and tablets; larger arm-based gestures are an option for interacting with touch-enabled all-in-one personal computers; and full-arm gestures are used to interact with the latest generation of gaming consoles. In the previous two chapters, I recommended using full-arm pointing-based gestures to interact with smart environments. Gestures and full-body interactions bring large easily-observable actions to general-purpose computers, and could thus be a solution to the problem of observability for collocated environments from domestic environments over interactive meeting rooms to industrial control room—they could be one way that designers help people maintain group awareness.

There is little information available, however, about whether gestural commands are in fact observable and interpretable, and what size of gesture is needed for an observer to notice the

gesture while carrying out other tasks. That is, how should gestures be designed to make possible the kind of group awareness that Norman described?

There are several factors that can influence people's ability to observe and identify the actions of an actor. The first one is gesture size, which is the physical size of the gesture motion. Figure 82 illustrates two gestures of different sizes in approximately the same scale. My hypothesis is that smaller gestures will be harder to observe and identify than larger ones.

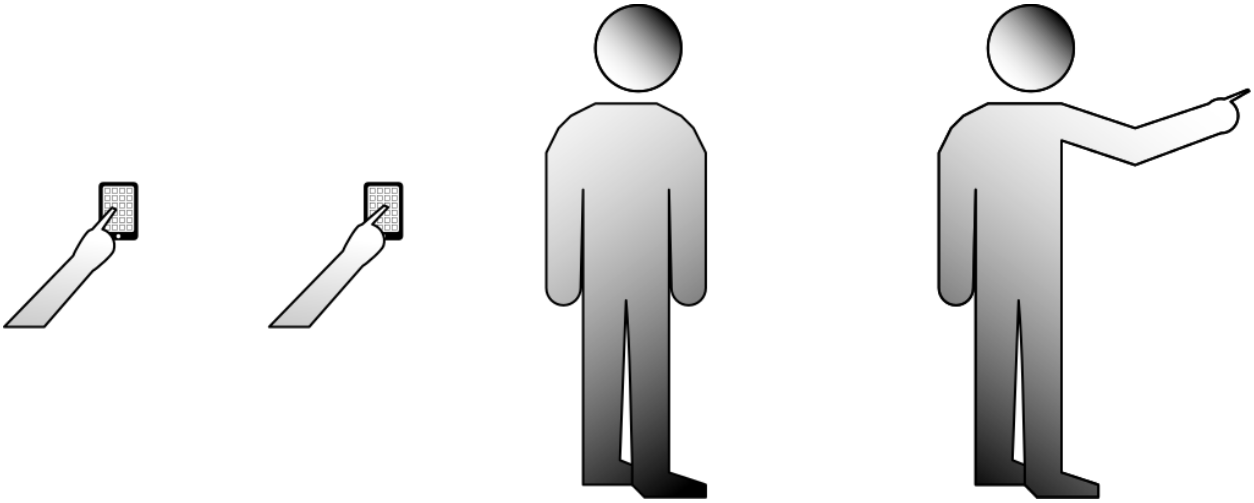


**Figure 82: Small gestures performed on a smart phone (two left); large full-arm gestures performed mid-air (two right)**

Another important factor is the mutual spatial orientation between people. Figure 83 shows two people (gray) observing someone performing an action (green). My hypothesis is that people facing toward an actor have less difficulties observing and identifying gestures than people who have the actor only within their outer field of view.

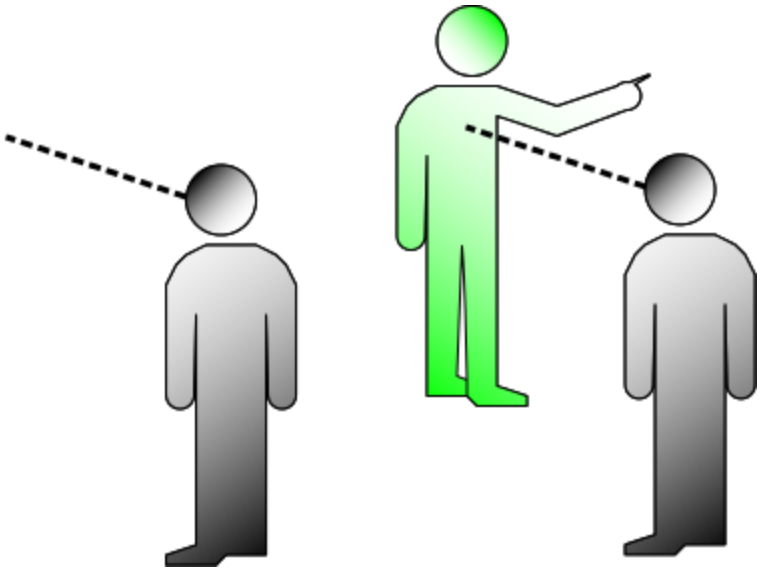


**Figure 83: Left person facing away from actor (green); right person facing toward actor; dotted lines indicate viewing direction.**

Last, the morphology of a gesture might play an important role in identifying them or, more precisely, in distinguishing between them. Figure 84 shows three types of gestures: a single tap, a double tap, and a swipe. My hypothesis is that gestures with similar morphology, e.g., a single and a double tap, are harder to identify than gestures with more different ones, e.g., single tap and swipe.

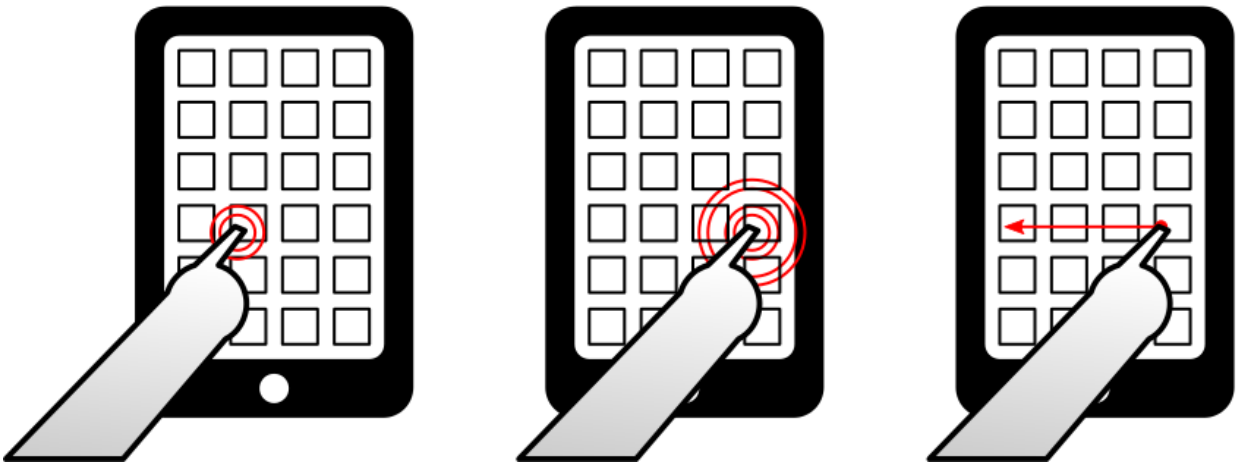


**Figure 84: Single tap on an icon (left), double tap (center), swiping across the screen (right)**

I carried out an experiment to answer this question. I looked at the following three issues in particular:

- How large is the effect of gesture size on the observability and identifiability of a gesture?
- What is the relationship between mutual orientation between actor and observer and observability and identifiability of a gesture?
- How does the gesture morphology influence observability and identifiability of a gesture?

Based on existing research in consequential communication (see 2.2.4), I formulated three hypotheses:

1. People can observe physically larger gestures more frequently than smaller ones.
2. People can identify physically larger gestures more accurately than smaller ones.
3. People can observe gestures more frequently and identify them more accurately when people are facing the actor.

## 7.2 Study Conditions

The following sections provide details on the types of gestures used in this study and the spatial relation between actor and observer.

### 7.2.1 Gesture Size and Morphology

I defined three gesture sizes: small touch-gestures performed on a 7″ hand-held tablet (see Figure 87); medium hover-gestures performed approximately 1 $cm$ above a 22″ horizontal screen (see Figure 86); and large full-arm pointing gestures (see Figure 85).

**Figure 85: Large gestures (point: left, high; point: front, high; point: right, high; point: left, low; point: front, low; point: right, low)**

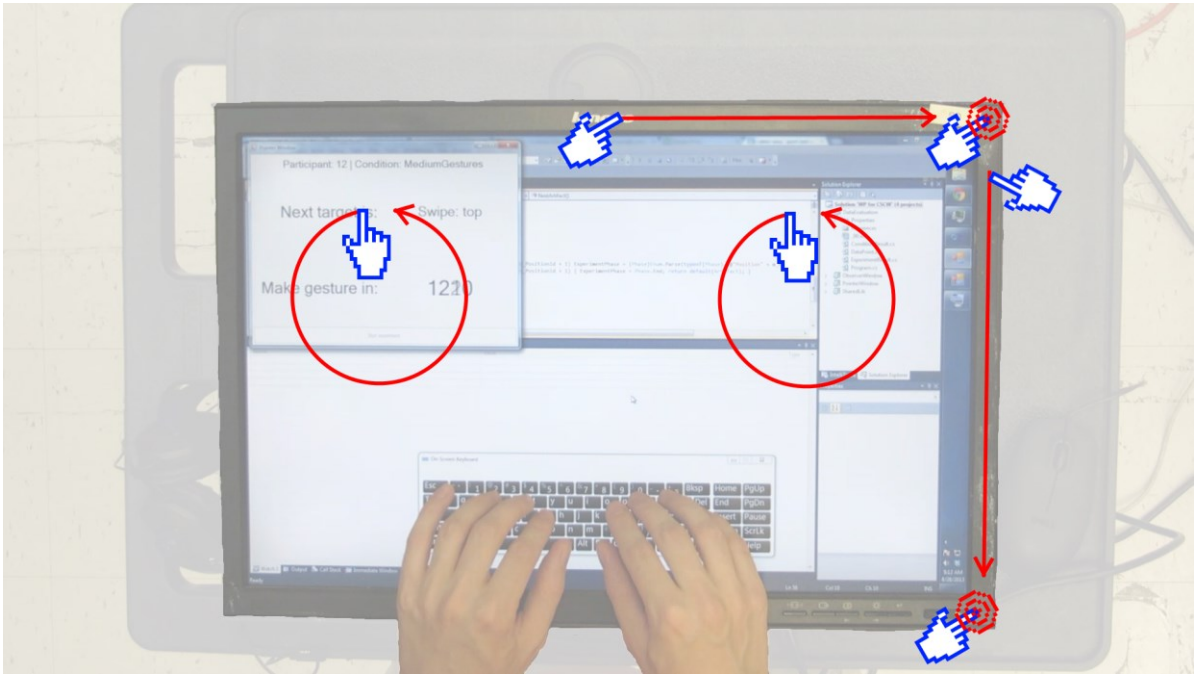



**Figure 86: Medium gestures (tap: top right corner, tap: bottom right corner, circle: left half, circle: right half, swipe: top edge, swipe: left edge); blue hand indicates starting point, red arrow trajectory of the gestures**

**Figure 87: Small gestures (tap: top left corner, tap: top right corner, circle: top half, circle: bottom half, swipe: left edge, swipe: right edge); blue hand indicates starting point, red arrow trajectory of the gestures**

For each of the gesture sizes, I created 6 different gestures (see Figure 85 to Figure 87 and Table 9). I chose a small gesture vocabulary in order to keep the recognition task simple, and to focus on my main interests of gesture observability and identifiability. For small and medium gestures, I chose two gesture types that can be found on most touch screens (tap and swipe) and one geometric gesture (circle). The large gestures were mid-air full-arm pointing gestures, similar to the ones used in *Room Pointing* and Ray-casting Air-pointing (Cockburn et al., 2011). For my system, I used six gestures that were arranged in front of the actor ($-90°$, $\pm 0°$, and $+90°$ horizontally, $-45°$ and $+45°$ vertically). During the experiment, I required participants to identify observed gestures only by their physical description, e.g. "Tap: top left corner" or "Point: left, high". I decided against asking for further interpretations of the observed and identified pointing gestures, e.g. "Tap: top left corner" → "Open browser" or "Point: left, high" → "Ceiling window" → "Open browser", because asking for interpretations would have introduced more possible error sources and, thus, confounded the results.

To gain quantifiable values for each of the gesture sizes, I measured magnitude and execution time of the actor's arm movement with an IR-based motion-tracking system. The actor performed each gesture 10 times while I captured his shoulder, elbow, wrist, and index finger movement. I then averaged the travelled distance and gesture time over all 10 trials. Naturally, the index finger travelled the longest distance: $\bar{x} = 0.46\ m$ (small gestures), $\bar{x} = 0.94\ m$ (medium gestures), and $\bar{x} = 1.65\ m$ (large gestures). Small gestures were performed in $\bar{x} = 1.9\ s$, medium gestures in $\bar{x} = 2.3\ s$, and large gestures in $\bar{x} = 1.7\ s$.

**Table 9: Gestures with mean magnitude and execution time**

| Small | Medium | Large |
|---|---|---|
| Tap: top left corner $(0.40\ m, 1.8\ s)$ | Tap: top right corner $(0.77\ m, 1.7\ s)$ | Point: left, high $(2.09\ m, 1.7\ s)$ |
| Tap: top right corner $(0.30\ m, 1.6\ s)$ | Tap: bottom right corner $(0.60\ m, 1.9\ s)$ | Point: front, high $(1.55\ m, 1.7\ s)$ |
| Circle: top half $(0.58\ m, 2.1\ s)$ | Circle: left half $(1.10m, 2.6\ s)$ | Point: right, high $(2.05\ m, 1.7s)$ |
| Circle: bottom half $(0.50\ m, 2.0\ s)$ | Circle: right half $(1.06\ m, 2.5\ s)$ | Point: left, low $(1.16\ m, 1.7\ s)$ |
| Swipe: left edge $(0.55\ m, 1.9\ s)$ | Swipe: top edge $(1.03\ m, 2.3\ s)$ | Point: front, low $(1.16\ m, 1.7\ s)$ |
| Swipe: right edge $(0.44\ m, 1.9\ s)$ | Swipe: left edge $(1.05\ m, 2.4\ s)$ | Point: right, low $(1.88\ m, 1.6\ s)$ |

### 7.2.2 Observer Location

Participants observed the actor from seven different locations (L1–L7), comprising of three positions arranged in a semicircle around the actor, and either two or three orientations at each position (facing the actor, or facing perpendicularly away). Figure 88 shows these locations. Six locations formed a symmetry: L1–L7, L2–L6, and L3–L5. The first two pairs, however, differ in a way that in L1 and L2 participants are behind the actor; in L6 and L7 they are in front.

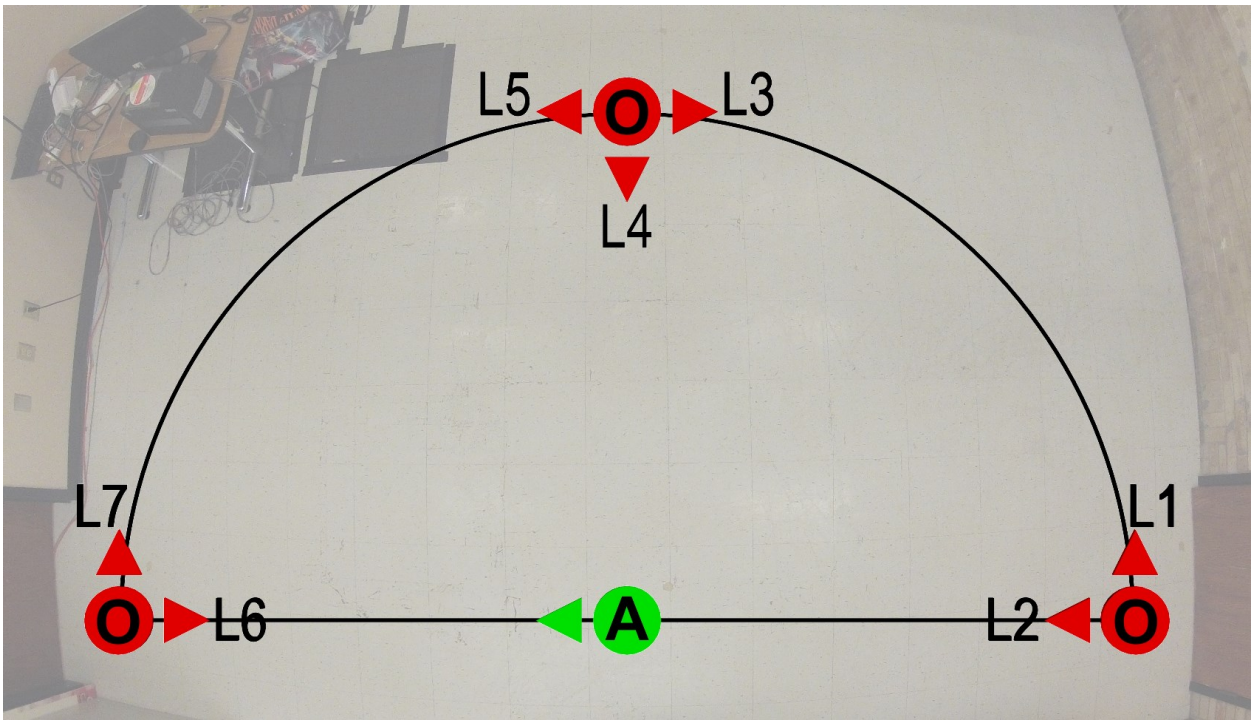



Figure 88: Observer locations (O) and actor location (A)

## 7.3 Experimental Setup

### 7.3.1 Study Design, Participants, and Apparatus

The study used a two-factor within-participant design with gesture size as a three-level factor (small, medium, large) and observer location as a seven-level factor (L1, …, L7). The order of gesture size was balanced using a Latin square, the order of observer location was randomized.

I recruited 18 participants (9 female, 9 male; ages $19-45$, $\bar{x}=29$ years) from a local university. These participants were all experienced with traditional computer systems ($\bar{x}=35\ h/wk$), and were all familiar with gestures on touch-based devices such as mobile phones and tablets. They received a $10 honorarium for participating in this one-hour-long study.

The study was carried out in a large laboratory (approximately $10\times10\ m^2$), in which I placed two moveable carts holding the study computers. The actor's cart held a 22″ monitor and remained stationary during the study. The observer's cart was moved to several different locations during the session (see Figure 88). It held a 7″ MiMo touch screen, on which the

primary task was displayed, and on which the observer indicated their observations and identifications of the actor's gestures.

### 7.3.2 Observer's Primary Task

In order to simulate a realistic work environment, I created an attention-demanding primary task for the observer to perform during the experiment. The task involved repeatedly selecting one of four possible buttons indicated by a written message displayed on the observer's display (displayed on a 7″ MiMo touch screen, see Figure 89). Participants were given a short period to complete the selection ($1.0\ s - 2.0\ s$, randomly chosen); if they did not finish their selection in time or made a wrong selection, the system would play a warning sound. After each correct selection, the system would wait $1.0\ s$ and then display another choice selection task.

### 7.3.3 Study Conditions and Procedure

After completing a demographics survey and being introduced to the system, participants completed 12 training trials. Participants were then moved to the starting location and asked to start the primary task; they were instructed to maintain awareness of the actor's activities, and report any gestures they observed using their interface (see Figure 89).

The actor then started performing typical tasks at his station, which acted as distractor tasks in between gestures that the observer had to report. The actor texted on the hand-held tablet (small gestures); he typed using the on-screen keyboard on the horizontal screen (medium gestures); and he fidgeted and moved objects around at the cart (large gestures). Within these typical activities, the actor performed a total of 12 gestures (each gesture twice, randomized order) per location. The actor's UI indicated when to perform the next gestures; the interval was randomly chosen (from $1.5\ s$ to $6.0\ s$). When participants noticed a gesture, they could pause the primary working task, and specify the gesture they just observed from the UI.

The actor performed a total of $(12 + 12 \times 7) \times 3 = 288$ gestures per participant. The observer's system recorded all gesture observations and identifications, and tracked the participant's performance on the primary task. After the experiment, participants filled out a basic demographic questionnaire, one NASA TLX form per gesture size, and one ranking questionnaire.

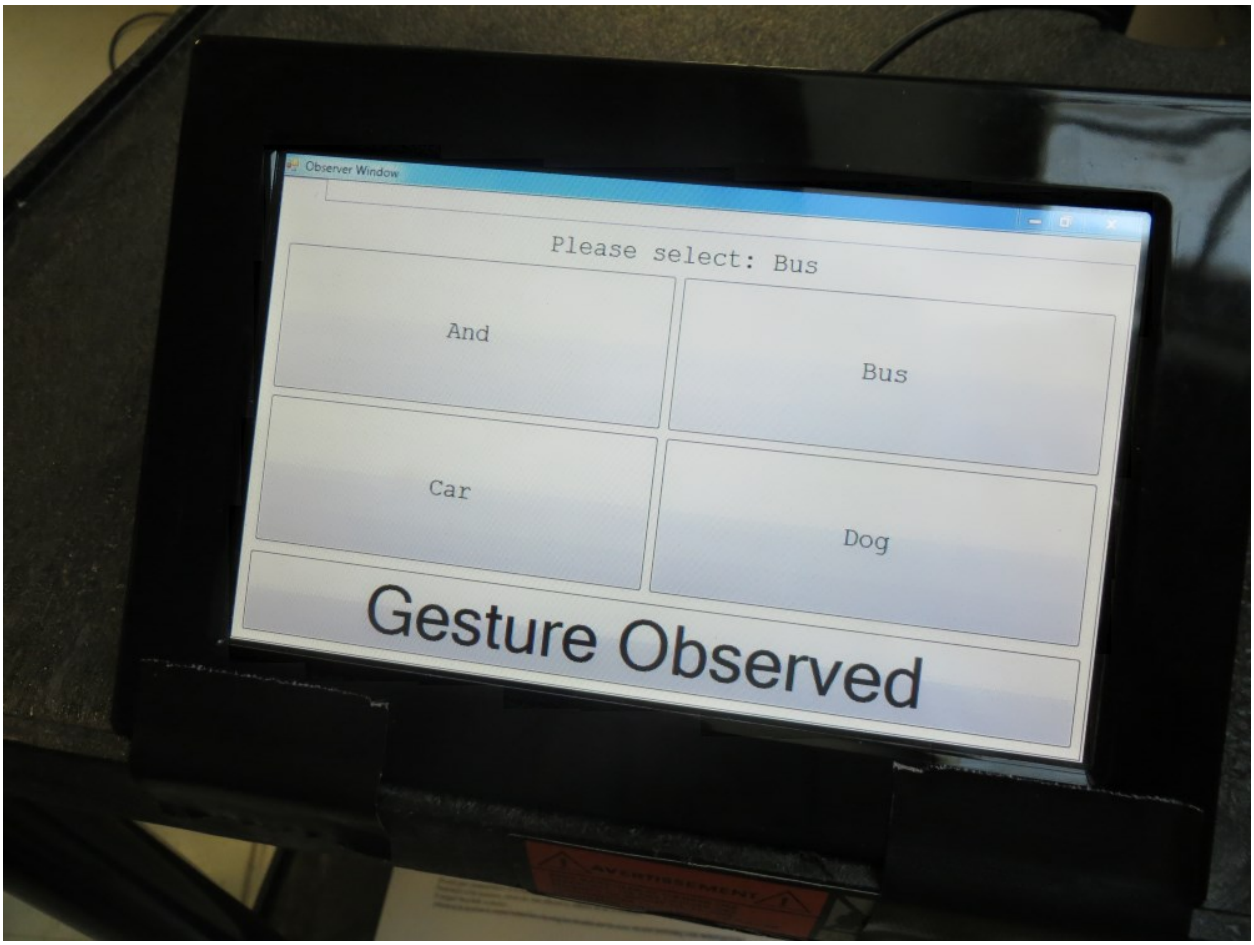



**Figure 89: User interface for the primary working task**

### 7.3.4 Data Analysis

I performed a univariate ANOVA to investigate the effect of gesture size and location on primary task performance measured as reaction rate. To determine the effect of the factors gesture size and location on observation and identification rates, I analyzed the trials in a 3×7 repeated-measures ANOVA. I carried out separate analyses of my dependent measures by gesture morphology (since morphologies were not the same across sizes) with a 3×6 RM-ANOVA. Last, I evaluated the TLX data using a repeated-measures ANOVA, and I analyzed the rank data using a Friedman test for k related samples. All post-hoc tests used Bonferroni corrections.

## 7.4 Results

### 7.4.1 Primary Task Performance

I did not find effects of Gesture size ($F(2,216) = 1.1, p > .1$) and Location ($F(6,216) = 1.8, p > .1$) on primary task performance. I found, however, a significant effect of Participant x Gesture size for 8 participants ($F(34,216) = 4.3, p < .01$). When looking more closely at this finding, I saw that all affected participants performed significantly worse with the first gesture size they saw during the experiment. I concluded that the training phase was too short for them to achieve their highest level of proficiency. Since I counter-balanced the order of gesture sizes between participants and therefore controlled for this factor, I felt confident that primary task performance was independent from gesture size and location. As a result, I omitted it from all further analyses.

### 7.4.2 Observation Rate and Identification Rate

Observation rate is the number of gesture observations made by a participant divided by the number of gestures performed by the actor. Identification rate is the number of gestures correctly identified by a participant divided by the number of observations.

Sphericity was violated for observation rate by both Gesture size and Location (Mauchly's test: $p = .00$), and for identification rate by Location (Mauchly's test: $p < .01$). For these analyses, I use Greenhouse-Geisser corrections.

### 7.4.3 Effects of Gesture Size

On average, participants showed the highest observation (see Figure 90 and Table 10) and identification rates (see Figure 91 and Table 10) with large gestures, followed by medium and small gestures.

**Table 10: Observation and identification rates per gesture size [%]**

| **Gesture Size** | Small | Medium | Large |
|---|---|---|---|
| **Observation rate: Mean ± Std. err.** | $74 \pm 5.0$ | $82 \pm 4.3$ | $83 \pm 3.0$ |
| **Identification rate: Mean ± Std. err.** | $69 \pm 3.8$ | $82 \pm 3.5$ | $92 \pm 2.0$ |

### *Observation Rate*

ANOVA showed a significant effect of Gesture size on Observation rate ($F(1.1,18.9) = 9.7, p < .01$). Follow-up analyses showed that medium and large gestures had a higher observation rate than small gestures ($p < .05$).

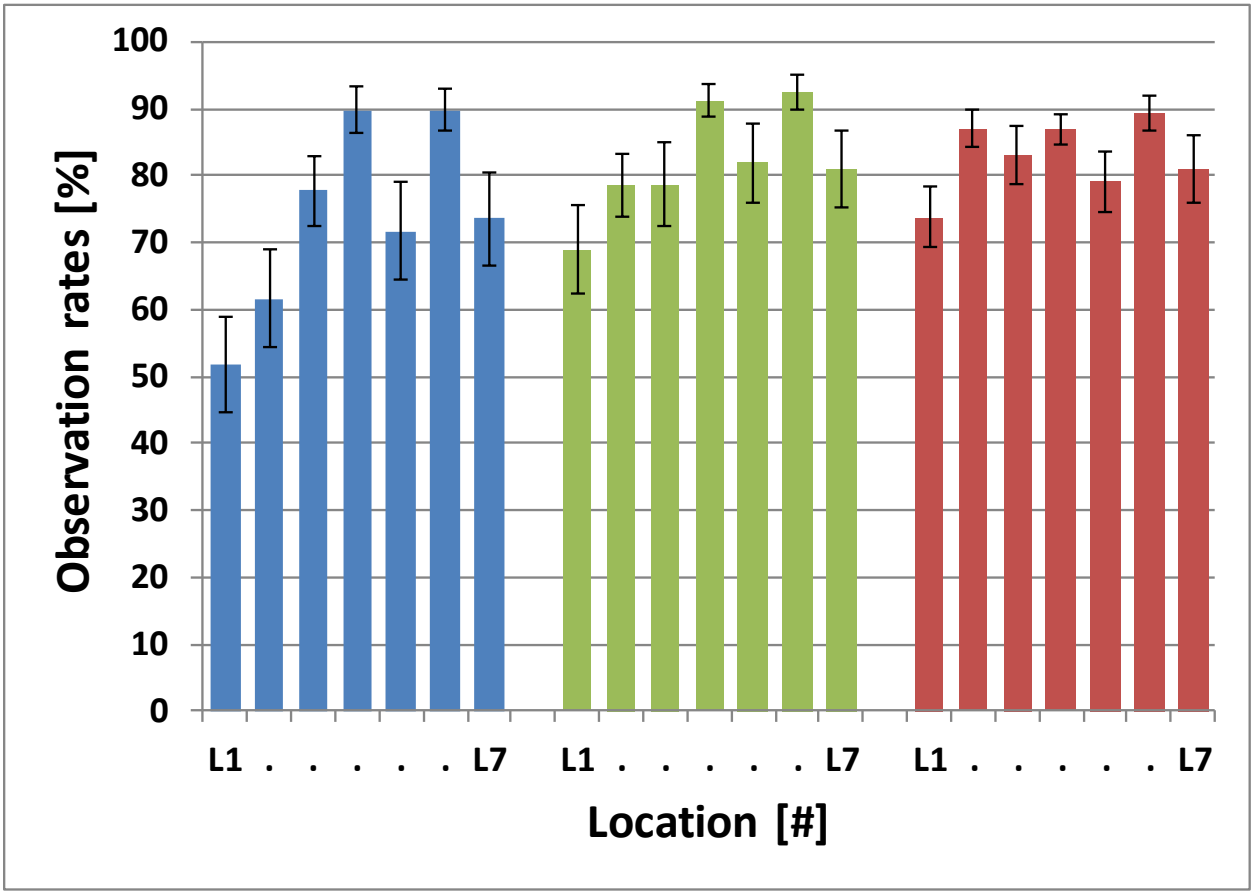



**Figure 90: Observation rates per gesture size**
**(small: blue / left; medium: green / center; large: red / right)**

### *Identification Rate*

ANOVA also showed a significant effect of Gesture size on Identification rate ($F(2,34) = 51.2, p = .00$). Follow-up analyses showed that all three gesture sizes were significantly different ($p < .01$).

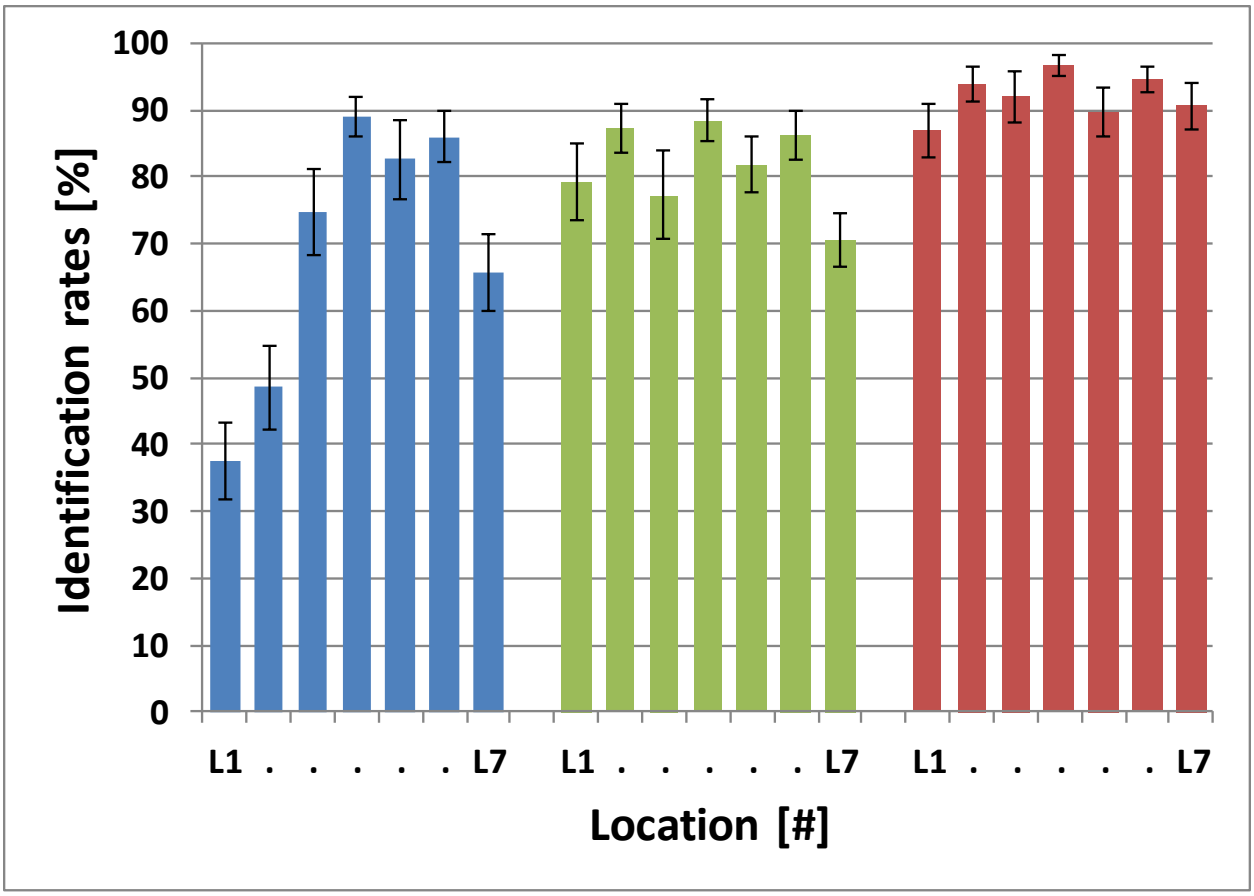



**Figure 91: Identification rates per gesture size**
**(small: blue / left; medium: green / center; large: red / right)**

### 7.4.4 Effects of Location

#### *Observation Rate*

ANOVA showed a significant effect of Location on Observation rate ($F(2.4,40.0) = 15.8, p = .00$). As shown in Table 11, the different locations were associated with a wide variety of observation rates: the highest at L6 and L4, and the lowest at L1 (see Figure 92 for a map of observation rates by location). Follow-up analyses showed that the observation rate at L1 was significantly worse than from all other locations ($p < .05$), and L6 had a higher observation rate than its symmetric counterpart L2 ($p < .01$).

### *Identification Rate*

ANOVA also showed a significant effect of Location on Identification rate ($F(3.2,53.7) = 13.4, p = .00$). As shown in Table 11 and Figure 93, participants had the highest identification rate from L4, followed by L6, and the worst observation rate from L1. The identification rate from L1 was significantly worse than from L3 through L6 (all $p < .05$), and L4 and L6 had significantly higher identification rates than L1, L2, and L7 (all $p < .01$)

**Table 11: Observation and identification rates per observer location [%]**

| **Location** | L1 | L2 | L3 | L4 | L5 | L6 | L7 |
|---|---|---|---|---|---|---|---|
| **Observation rate: Mean ± Std. err.** | 65 ± 5.4 | 76 ± 4.2 | 80 ± 4.5 | 89 ± 2.1 | 78 ± 5.7 | 91 ± 2.5 | 79 ± 5.5 |
| **Identification rate: Mean ± Std. err.** | 68 ± 4.5 | 77 ± 3.4 | 91 ± 5.0 | 91 ± 1.8 | 85 ± 3.7 | 89 ± 2.5 | 76 ± 3.3 |

## 7.4.5 Gesture Size x Location Interaction

### *Observation Rate*

ANOVA showed a significant interaction between Gesture size and Location ($F(5.7,97.4) = 4.1, p < .01$) for Observation rate. As shown in Figure 90, small gestures were significantly better observed from L4 and L6 (both $\bar{x} = 0.90$) than from L1 ($\bar{x} = 0.52$) and L2 ($\bar{x} = 0.62$) (all $p < .01$). Observation rate from L1 was significantly worse than from all other locations except L2 (all $p < .05$). As expected, mean differences were high between symmetric locations L1–L7 ($.22$) and L2–L6 ($.28$) and low between L3–L5 ($.06$) and L4–L6 ($.00$).

Medium gestures were best observed from L6 ($\bar{x} = 0.93$) and L4 ($\bar{x} = 0.91$) and worst observed from L1 ($\bar{x} = 0.69$). Observation rates from L1 were significantly worse than from L3 through L6 (all $p < .05$) and from L2 worse than from L6 and L4 (both $p < .05$). Compared to small gestures, observation rate in L2 improved close to average ($\bar{x} = 0.79$). As expected, mean differences became lower between symmetric locations L1–L7 ($.12$) and L2–L6 ($.14$) and stayed low between L3–L5 ($.03$) and L4–L6 ($.01$).

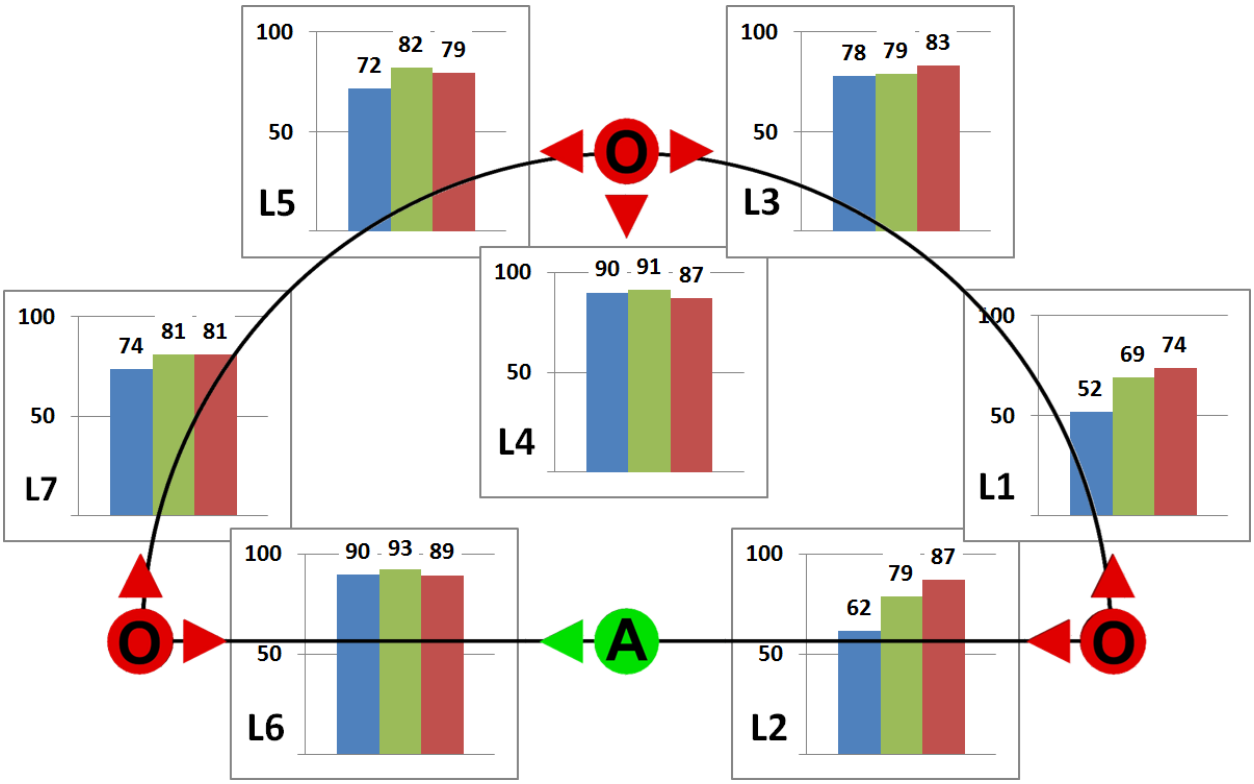



**Figure 92: Observation rates per location**
**(small: blue / left; medium: green / center; large: red / right)**

Large gestures were best observed from L6 ($\bar{x} = 0.89$) and L2 and L4 (both $\bar{x} = 0.87$) and worst observed from L1 ($\bar{x} = 0.74$). The only significant difference appeared between L6 and L1 ($p < .05$). Overall, differences between symmetric locations are for amongst large gestures: L1–L7 ($.07$) and L2–L6 ($.02$), L3–L5 ($.04$) and L4–L6 ($.02$).

At L1 and L2, participants showed significantly lower observation rates with small gestures than with medium and large gestures (all $p < .05$).

### *Identification Rate*

ANOVA also showed a significant Gesture size × Location interaction ($(F4.5,76.7) = 12.5, p = .00$) for identification rate. As shown in Figure 91, small gestures were significantly better identified from L4 ($\bar{x} = 0.89$) and L6 ($\bar{x} = 0.86$) than from L1 ($\bar{x} = 0.37$), L2 ($\bar{x} = 0.49$), and L7 ($\bar{x} = 0.66$) (all $p = .00$). Identification rates from L1 and L2 were significantly worse than

from all other locations (all $p < .05$). As expected, mean differences were high between symmetric locations L1–L7 (.28) and L2–L6 (.37) and low between L3–L5 (.08) and L4–L6 (.03).

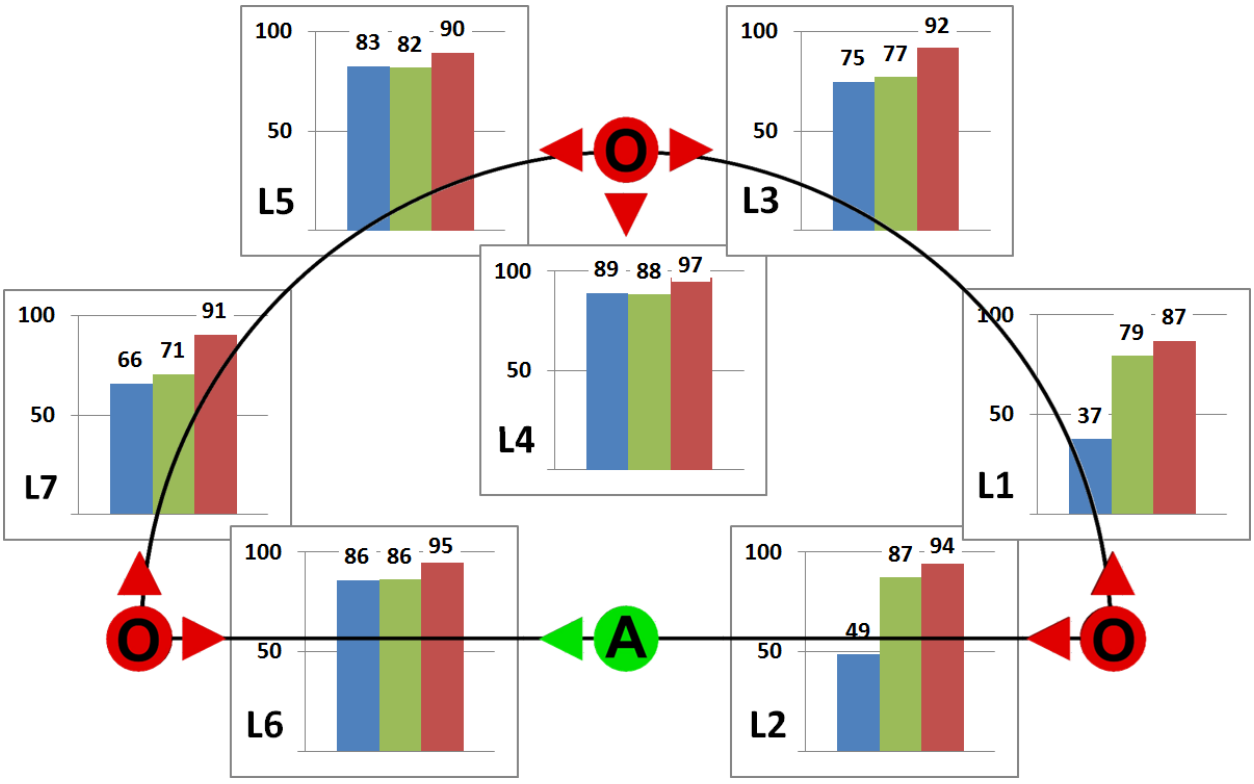



**Figure 93: Identification rates per location**
**(small: blue / left; medium: green / center; large: red / right)**

Medium gestures were best identified from L4 ($\bar{x} = 0.88$) and L6 ($\bar{x} = 0.86$) and worst identified from L7 ($\bar{x} = 0.71$). Identification rate from L7 was significantly worse than from L2, L4, and L6 ($p < .05$). As expected, mean differences became low between symmetric locations L1–L7 (.09) and L2–L6 (.01) and stayed low between L3–L5 (.05) and L4–L6 (.02).

Large gestures were best identified from L4 ($\bar{x} = 0.97$), L6 ($\bar{x} = 0.95$) and L2 ($\bar{x} = 0.94$) and worst identified from L1 ($\bar{x} = 0.87$). There were no significant differences between locations. Overall, mean differences between symmetric locations are very similar with large gestures: L1–L7 (.04) and L2–L6 (.01), L3–L5 (.03) and L4–L6 (.02).

At L1 and L2, participants showed significantly lower identification rates with small gestures than with medium and large gestures (all four $p = .00$); at L3 and L7, participants showed significantly higher identification rates with large gestures than with small and medium gestures (all $p < .05$).

### 7.4.6 Effects of Gesture Morphology

I analyzed gesture morphology separately within each gesture size (since they were different across sizes). Since sphericity was violated for all measures (Mauchly's test: all $p < .05$), I use Greenhouse-Geisser corrections.

For small gestures, ANOVA showed a significant effect of Gesture morphology on Observation rate ($F(3.6,61.1) = 5.7, p < .01$) and on Identification rate ($F(3.0,51.4) = 6.1, p < .01$); for medium gestures, ANOVA showed a significant effect of Gesture morphology on Observation rate ($F(3.5,60.0) = 5.7, p < .01$) and on Identification rate ($F(2.7,46.5) = 5.3, p < .01$); for large gestures, ANOVA showed a significant effect of Gesture morphology on Observation rate ($F(2.4,61.0) = 17.1, p = .00$), but not Identification rate.

### *Observation Rate*

I found that participants observed the small gesture circle: top significantly more often than the gestures tap: top left, tap: top right, and circle: bottom (all $p < .05$). For medium gestures, participants observed tap: bottom right significantly less often than the gestures circle: left, circle: right, tap: top right, and swipe: top (all $p < .05$). Among large gestures, point: left, low had a significantly lower observation rate than any other large gestures (all $p < .05$), and was the least-observed gesture at any size.

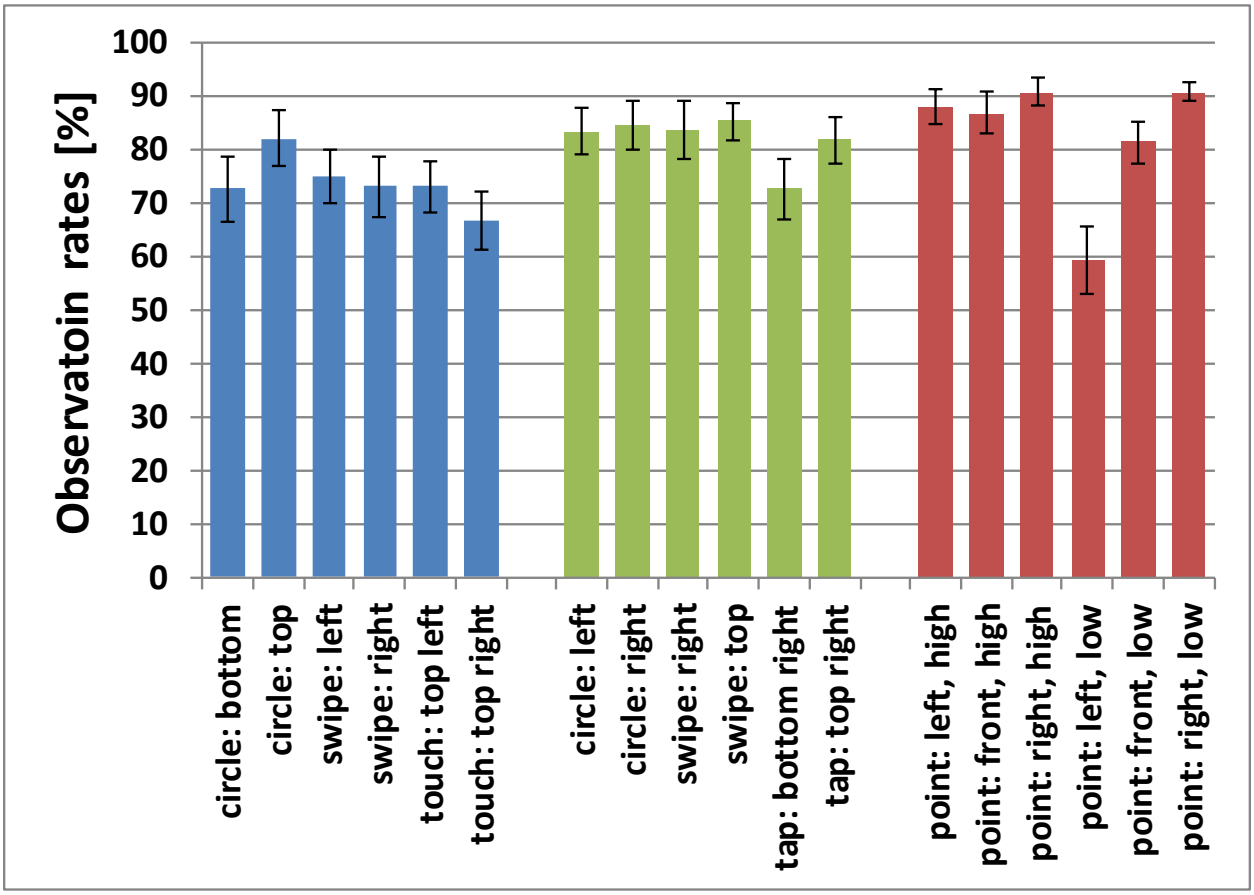



**Figure 94: Observation rates per gestures**
**(small: blue / left; medium: green / center; large: red / right)**

### *Identification Rate*

Participants showed a significantly higher identification rate for the small gesture swipe: right than for the gestures swipe: left and tap: top right (all $p < .05$). For medium gestures, participants identified swipe: top significantly less often than all gestures except tap: bottom right ($p < .05$). For large gestures, there were no significant differences.

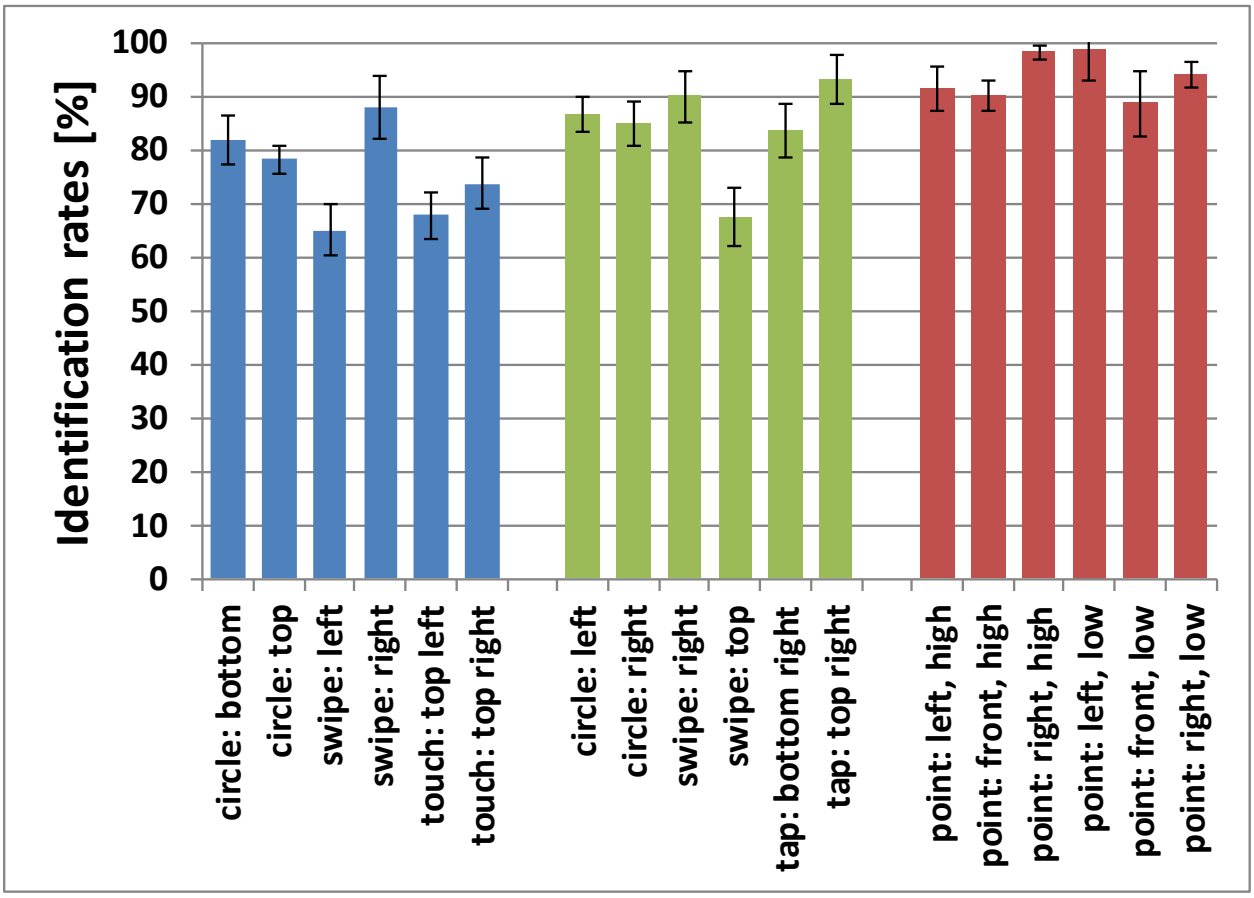



**Figure 95: Identification rates per gestures**
**(small: blue / left; medium: green / center; large: red / right)**

## 7.4.7 Subjective Measures

Participants rated their experience using the NASA TLX questionnaire. Overall, participants felt that larger gestures were less effort and less frustration than smaller gestures.

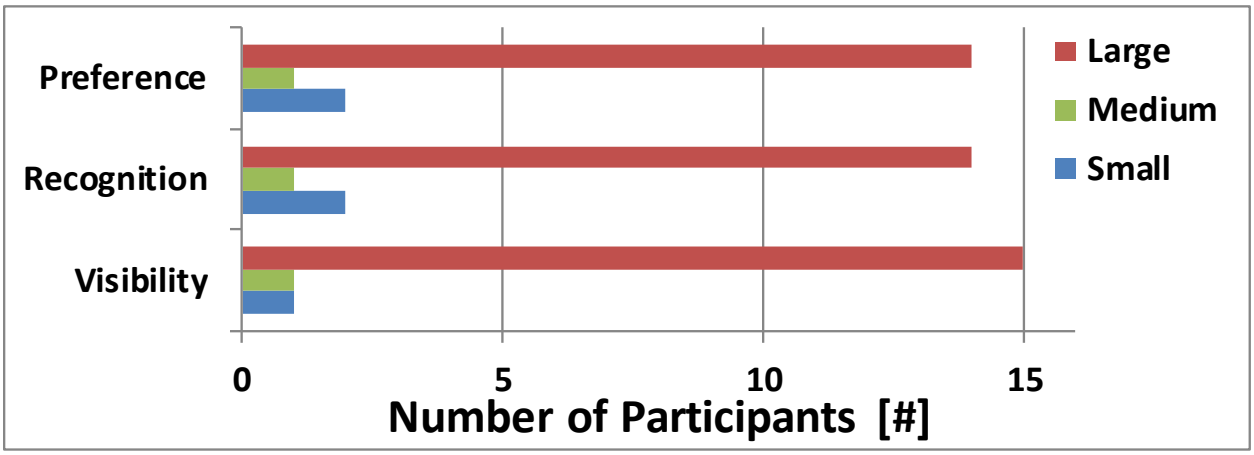



Figure 96: Participant preference rating

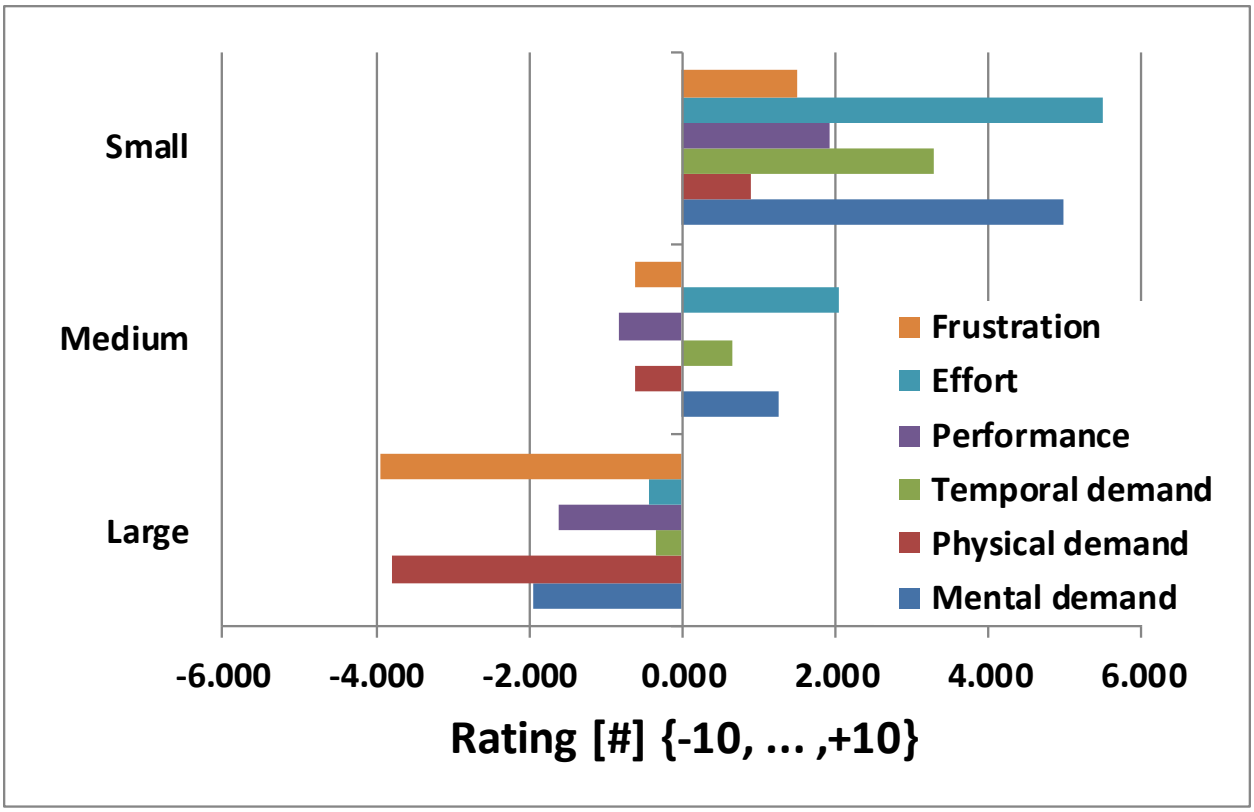



Figure 97: NASA TLX results

I found a significant difference in mental demand between all three gesture sizes (all $p < .01$). For physical demand and frustration, there were significant differences between large gestures and small and medium gestures (both $p < .05$). Finally, participants rated small gestures as more effortful than medium and large ones (both $p < .01$).

I also asked participants to rank the different gesture sizes in terms of perceived visibility, recognition accuracy, and their preference to work with (Figure 97). (I discarded the data from one participant because the questionnaire was not filled out correctly.) A significant majority of participants ranked large gestures most visible ($15/17$: $\chi^2(2,17) = 21.5, p = .00$) and most recognizable ($15/17$: $\chi^2(2,17) = 19.9, p = .00$). Overall, 14 of 17 participants preferred to work with large gestures over small and medium ones ($\chi^2(2,17) = 17.3, p = .00$).

## 7.5 Discussion

In this discussion, I first explain how my results confirm my hypotheses and discuss some additional insights I gained from analyzing my results. Then, I come back to my premise and lay out how my findings support Norman's idea of "big controls and big actions" (see 2.2.4). I describe some use cases, mention potential directions for future work, and address issues that come with the use of big gestures. Finally, I list the limitations of my work.

### 7.5.1 Review of the Main Hypotheses

At the beginning of this chapter, I postulated three hypotheses:

1. People can observe physically larger gestures more frequently than smaller ones
2. People can identify physically larger gestures more accurately than smaller ones
3. People can observe and identify gestures better when facing the actor

#### *Larger Gestures are more Frequently Observed*

As predicted, participants showed significantly higher observation rates with large and medium gestures than with small gestures, and higher observation rates with large gestures than with medium gestures. While this result is true on the (categorical) gesture-size scale (small—medium—large), I also found a similar pattern when looking at the (continuous) gesture magnitude. Figure 98 illustrates the logarithmic relationship between gesture magnitude and observation rate ($F(1,15) = 119.0, p = .00, R^2 = .89$). However, my regression analysis

revealed that one large gesture ("point: left, low") was a residual outlier. For the curve fit, I removed this outlier (case-wise analysis with 3$\sigma$ cutoff); I talk about this case later in the discussion. I want to emphasize that the logarithmic relationship continues across different gesture sizes and morphologies (2D touch and hover gestures as well as 3D pointing gestures).

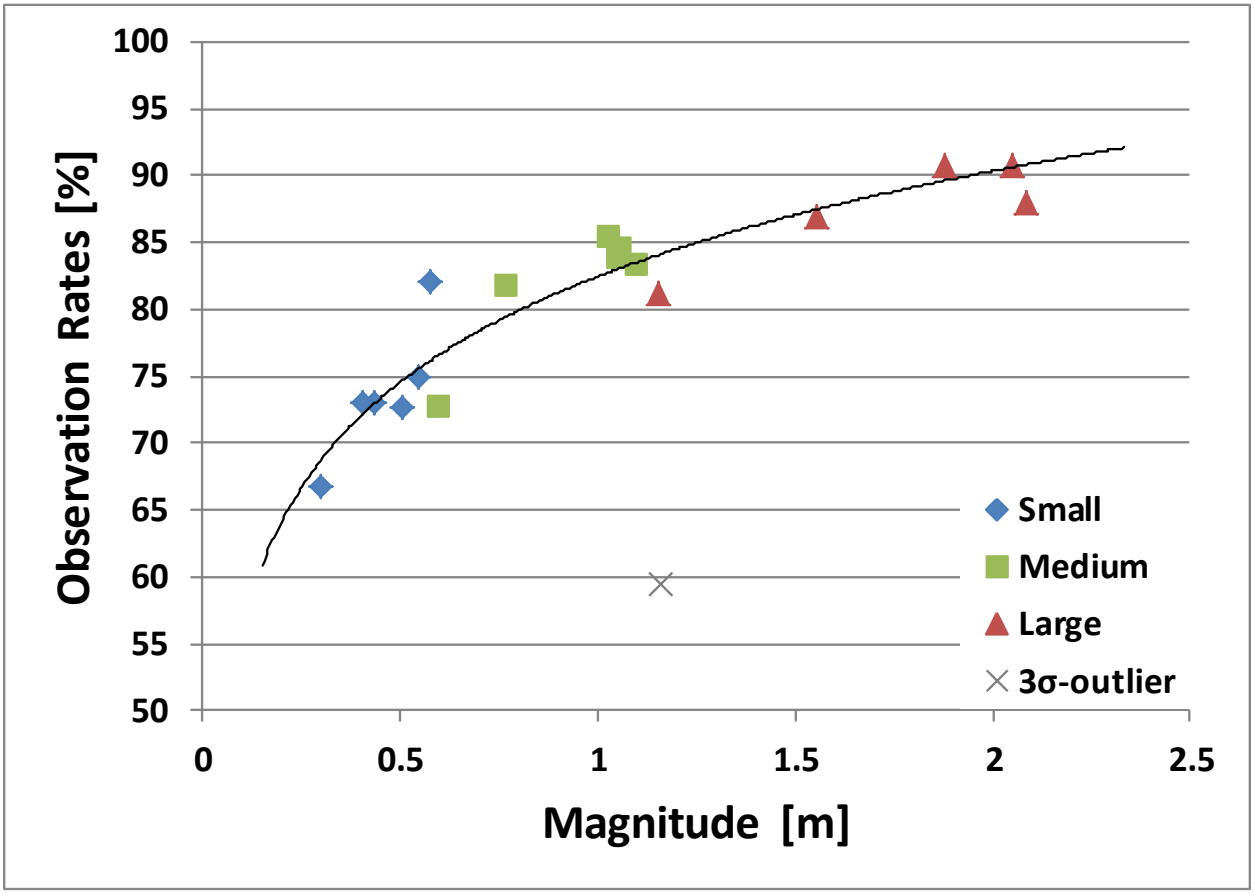



Figure 98: Observation rate per gesture magnitude

### *Larger Gestures are more Accurately Identified*

Participants showed significantly better performance with large gestures than with medium and significantly better performance with medium than with small gestures. The overall identification rate of larger gestures is better than that of smaller gestures; even when observed, larger gestures are easier to identify than smaller gestures.

I found a logarithmic relationship between magnitude and identification rate, similar to the one between magnitude and observation rate ($F(1,15) = 19.3, p < .01, R^2 = .56$). Not surprisingly, the effect is smaller because there are other factors that affect identification rate. Again, my regression analysis revealed, that one gesture (medium size, "swipe top") was a residual outlier. For the curve fit, I removed this outlier (case-wise analysis with $2\sigma$ cutoff); I will come back to this particular case later in the discussion.

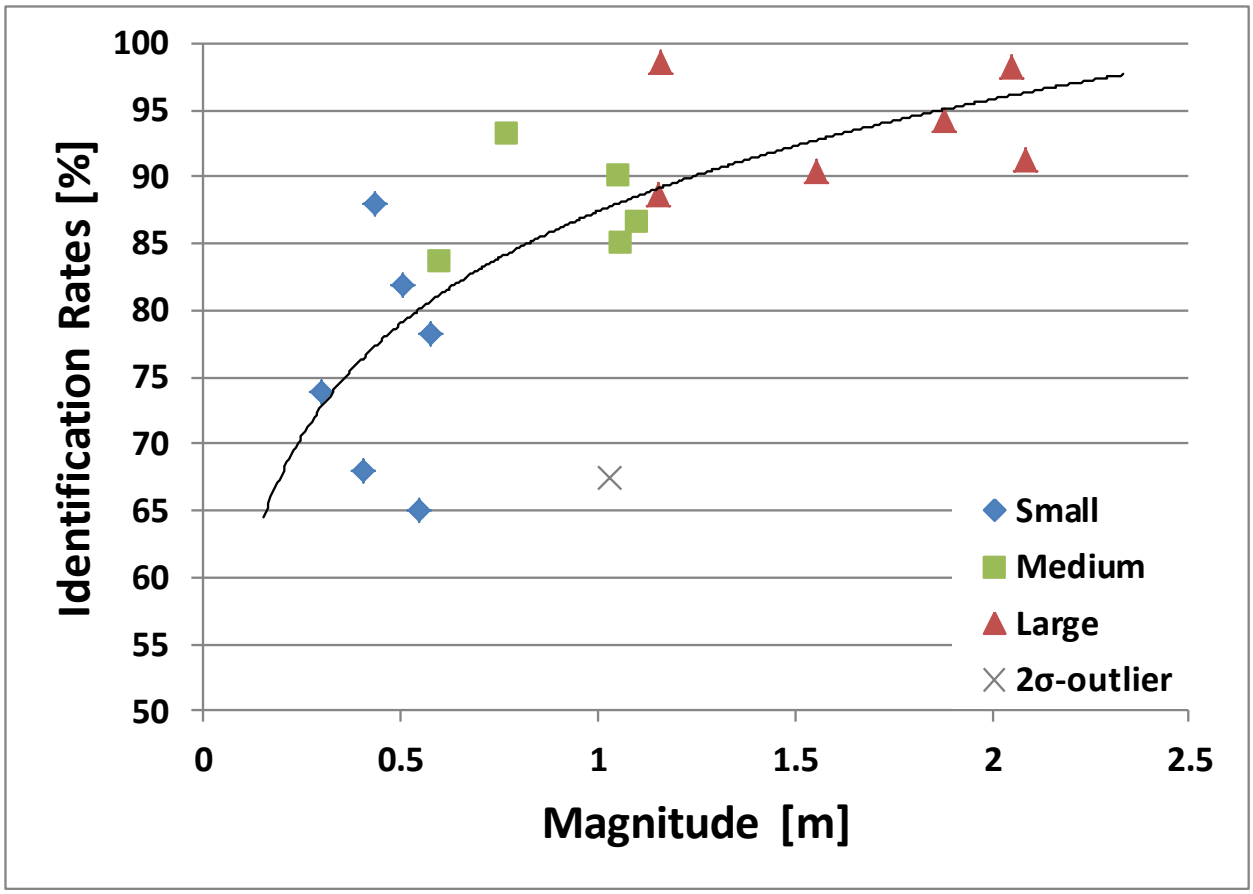



Figure 99: Identification rate per gesture magnitude

### *Facing the Actor Increases Gesture Observability and Identifiability*

In locations L2, L4, and L6, participants were facing the actor, in locations L1, L3, L5, and L7, they were perpendicularly seated to the actor. When pairwise comparing L1–L2, L3–L4, L5–L4, and L6–L7, I found that participants performed on average better when facing the actor. However, most of these comparisons showed no significant difference. While these results might

sound surprising, they are in accordance with theory. Vision research has shown that human response to rapidly moving targets is almost invariant with its location in the field of vision (Tynan and Sekuler, 1982).

### 7.5.2 Additional Findings and Research Questions

#### *Are all Gestures of One Size Equally Easy to Observe?*

For small gestures, I found that "circle: top" was the easiest gesture to observe, significantly easier than both "taps" and "circle bottom" (all $p < .05$). This was most likely because it had the longest execution time ($2.1\ s$) and largest magnitude ($0.58\ m$) among all small gestures.

For medium gestures, I found that "tap: bottom right" was significantly harder to observe than any other gesture except "swipe: right" (all $p < .05$). Contributing factors were its low execution time (second lowest in its category: $1.9\ s$) and its small magnitude (smallest in its category: $0.60\ m$).

For large gestures, I found that participants showed a significantly lower observation rate with "point: left, low" than with any other gesture (all $p < .05$). As before, I assume that mostly execution time ($1.7s$) and lack of magnitude ($1.16\ m$) are responsible for this effect. In addition, the gesture was performed very close to the body, which made it more difficult to spot than other large gestures, which were all performed away from the actor's body.

#### *Are all Gestures of One Size Equally Easy to Identify?*

For small and large gestures, I found no gesture that was consistently better or worse than the other ones.

For medium gestures, however, I found that participants performed significantly worse with "swipe: top" than with any other gesture except "tap: bottom right"(all $p < .05$). A detailed analysis showed that I can attribute more than half of the errors to confusing this gesture with the gesture "tap: top right". These two seemingly different gestures share a similar post-stroke hold and retraction phase. Apparently, participants oftentimes required the preparation and stroke phase of the actor's gesture to shift their attention from their primary working task to the perception phase of consequential communication. To make gestures more distinguishable, I therefore recommend avoiding gestures that end with similar strokes and the same post-stroke

hold and retraction. For example, the small-gesture swipes were rarely confused with the small-gesture taps.

#### *Are Large Gestures Generally Easy to Observe and Identify?*

Ironically, the least likely observed gesture in my study was a large one. A good strategy to make large gestures visible is to make them lead away from the actor's body.

#### *Did the Labels "Left" and "Right" Confuse Participants?*

All directions in gesture descriptions were meant to be relative to the actor. As a result, there was a danger that participants confused left and right and top and bottom when they were in front of the actor (e.g., his "left" became their "right"). I analyzed all errors in conditions L6 and L7; no participant systematically confused any of these labels.

#### *Are Larger Gestures less Affected by Occlusion?*

In locations L1 and L2, gestures were occluded by the actor's body. A comparison of symmetrical pairs L1–L7 and L2–L6 therefore shows how much occlusion affected participants' observation rate. My results showed that the mean differences in observation rate between L1 and L7 and between L2 and L6 decreased with increasing gesture size. This implies that small gestures suffer strongly from occlusion and that this effect diminishes with increased gesture size. With an unobstructed view to the actor, gesture size does not affect performance. However, in multi-display environments where people move around freely, it is likely that occlusion will occur; in this case, larger gestures can enable higher group awareness.

Identification rates of all gestures were affected in similar ways by occlusion than observation rate. For medium gestures, differences in identification rates between L1 and L7 and between L2 and L6 were smaller than these differences in observation rate. I suspect that the location of the gestures on the 22" screen were responsible for this effect: five out of six gestures were performed close to the right edge of the screen, so observers were able to catch a glimpse of these gestures around the right side of the actor's body.

### 7.5.3 "Big Controls and Big Actions"

Norman's original idea was that big controls and big actions create awareness. My results showed that gestures, independently from their size, are indeed observable and can therefore improve group awareness: people know that something has happened. When looking at

identification rate, I can also give an initial estimation for the next step toward group awareness, knowing what exactly has happened. My results indicate that people can distinguish between at least six gestures. I also showed that identification rate depends on more factors than observation rate. A more thorough investigation of these factors could give more insights about potential limitations, such as upper limits of an alphabet of discernible gesture, as well as guidelines for designing distinguishable gestures. Another important issue is finding gesture sets with different levels of observability, so that interaction designers can select a gesture that matches an action's desired publicity.

There are many cases in which people would want to make their actions public. In domestic environments, people can use public gestures to share information about their activities and intentions with others. Public gestures can also be part of, for example, co-located multiplayer games where the group should be aware of certain actions. Likewise, there are many cases in which people want to keep actions private or do not want to distract others. As said before, my findings show that people can control the publicity or privacy of their actions through gesture size.

There are, however, some disadvantages to large gestures. For example, they require more physical effort, and there are some socio-cultural restrictions to the use of big gestures. Again, I assume that large gestures will mostly be used in domestic or group environments, where each member accepts and understand large gestures in the context of their activity.

### 7.5.4 Limitations of this Study

There are a couple of limitations to my study. While I selected gesture sizes to reflect a broad variety of gestural interfaces, I only used a typical set of gestures within each size and not a broad variety of all possible gestures. This allowed me to only give an initial assessment and lower boundary about identification rates, leaving a more systematic approach to future work. Common contextual and semantic knowledge, for example, can increase identification rates. In addition, my study took place in a controlled laboratory environment.

## 7.6 Conclusion

Awareness of ones interaction with the environment is important for fostering communication and collaboration between co-located people. While creating awareness works well with in-place

interfaces and consequential communication, it might become difficult with navigation-based interactions, such as smart phones and tablets. In this chapter, I demonstrated that gestural interaction techniques can be used for creating visible HEI, thus laying the groundwork for providing consequential communication to co-located people. I measured observation and identification rates of different gestures and showed that even small gestures are visible and could create consequential communication. However, larger gestures are more easily observable, mainly due to a reduced effect from occlusion. In addition, increasing size makes gestures more easily identifiable. The use of mid-air full-arm pointing gestures in room-based interaction can help awareness creating through consequential communication.

# Chapter 8 General Discussion

In this section, I provide a summary of the main findings of my dissertation, discuss how the results from the three user studies relate to my claims and hypotheses about room-based interaction

## 8.1 Summary of Primary Findings

In the introduction, I argued that room-based interaction has three advantages over existing techniques for HEI: it allows for better performance, it offers device-free interaction, and it increases publicity of interactions (see 1.1). From this, I postulated the following three main claims:

1. Room-based interaction allows for faster interaction than navigation-based interfaces.
2. With sufficient training people might be able to use room-based interaction hands-, eyes-, and (system) feedback-free.
3. Room-based interaction allows for publicly visible interactions with smart environments.

I will now revisit these three claims and discuss their validity in view of the three user studies presented in my dissertation.

### 8.1.1 Selection Speed in Room-based Interaction

In the introduction, I argued that room-based interaction has inherent performance advantages over all in-place interfaces and performance advantages over navigation-based interfaces when the transition between primary task and HEI-task is costly (see 1.1.1). I did not verify this claim as simple reasoning should sufficiently show the performance advantage of room-based interaction in these two cases.

I my first study I investigated how performance of room-based interaction compares to navigation-based interfaces under optimal conditions for navigation-based interfaces (e.g., interaction device already at hand). In particular, I showed how much three different factors influence selection speed in HEI: organization of storage space (flat versus hierarchical), selection mechanism (touch versus pointing), and proxy type (on-screen icons versus real-world objects).

The results indicate that the organization of the storage space is decisive for the speed of an interaction technique: the three selection techniques that used a flat input space were significantly faster than the one that used a linear one (see 5.4.1). This result matches the assessment in my conceptual framework, which predicts that the time spent on finding the proxy icon within the input space depends on the structure of the input space (see 3.2.1). This finding confirms that room-based interaction can allow for faster interaction as soon as touch-based techniques have to use non-flat input space. With current touch devices, the transition from flat to hierarchical storage space normally occurs when the storage space has to hold more icons than fit on the screen (20∼25 for smart phones, 25∼45 for tablets). With room-based interaction, people can store a substantially larger number of items in the environment while retaining reasonably large proxy zones, e.g., 110 proxy zones with 20° diameter each (see 8.3.1).

Furthermore, the results show that using mid-air full-arm pointing gestures as selection mechanism and real-world objects as selection proxies in room-based interaction does not improve people's performance compared to navigation-based interfaces. On the contrary, selection speed with *Room Pointing* was slower than with *Screen Pointing*. This result might not surprise as the mid-air full-arm pointing gestures used in room-based interaction are physically larger and thus slower than the forearm-motions used in *Screen Pointing*. In addition, the lack of system-feedback results in overall lower selection accuracy for room-based interaction (see 5.4.2). This result is of course expected since system feedback is an important mechanism for error avoidance. The higher error rate is, however, not a disadvantage of room-based interaction: with room-based interaction, people have the choice whether to use system feedback or not, while navigation-based interaction inherently requires system feedback. When people use room-based interaction with system feedback, their selection accuracy is as high as with any other navigation-based interaction technique (see 5.4.2), but they can decide to trade-off selection accuracy for system-feedback-free (= eyes-free) interaction if the situation requires it.

In summary, my results show the performance advantages of room-based interaction and outline scenarios in which interaction designers should prefer room-based interaction. This includes all situations in which users want to perform device-free and eyes-free interactions, as this is a feature navigation-based interaction cannot offer, and scenarios in which people need fast access to a number of digital artifacts large enough so that navigation-based interaction has to switch to

a hierarchically organized storage space. In these scenarios, room-based interaction solves the problem of not having a fast interaction technique for HEI.

### 8.1.2 Interaction Devices and Feedback in Room-based Interaction

In the introduction, I argued that HEI techniques should provide the opportunity for device-, system-feedback-, and eyes-free interactions. Having access to techniques with different operation modalities and feedback channels would allow people to tailor their interaction to the situation, for example, device-free when people do not have their hands free for interaction or eyes-free when they do not want to shift visual focus.

#### *Interaction Devices and Device-free Interaction*

An inherent advantage of using mid-air full-arm pointing gestures, for example *Room Pointing*, *RCAP*, or *Screen Pointing*, is that people do not have to hold or touch any interaction device. This frees up people's hands for other tasks, which can be particularly helpful for HEI as many primary tasks in people's daily life require the using their hands. The results of my first study, where I compared touch- and pointing-based input, showed that people were able to use *Room Pointing*, *Screen Pointing*, and *Touch Flat* equally accurately and quickly as long as system feedback was provided. This shows, people can use pointing gestures as input mechanism as accurately and quickly as touch-based interfaces while having the advantage of not having to touch or hold any interaction device.

#### *Feedback Channels*

The results of my first study (see 5.4.2) showed that people's selection accuracy dropped once they did not receive any system feedback about their currently selected target. This indicates that the type of feedback plays an important role in selection accuracy of room-based interaction. In the context of y dissertation, I differentiate three types of feedback that people can receive: intrinsic feedback, which is naturally generated within the human body (e.g., proprioception, see 2.4.2); extrinsic feedback, which is naturally generated outside the human body (e.g., visual feedback (2.4.2); and system feedback, which is artificially generated by the digital system (e.g., mouse cursor and currently selected target in the first study, see 5.3.3). Generally, it is one of the basic principles for system design to give people feedback about the state of the system, mostly as a visual cue (Shneiderman, 1997). Evaluating system feedback during interaction, however, is an additional task and competes with the actual working task for people's cognitive capacity (see

2.2.4): having too much system feedback negatively affects people's task performance. Reducing the reliance on feedback, i.e. the automation of task execution, is therefore an important method for decreasing cognitive load and increasing efficiency of many routine tasks in people's daily life. While this task automation lies at the core of motor skill learning (see 2.5.4), it also applies to other types of procedural memory, e.g., always putting the keys on a particular place when entering the house (simple conditioning), and semantic memory, e.g., remembering the usual location of the keys (spatial memory, see 2.5.3). Unlike traditional touch-based interaction, room-based interaction does not require system feedback and thus allows people to decide what types of feedback to use: intrinsic feedback only, i.e. eyes-free interaction; extrinsic feedback, i.e. system-feedback-free interaction; or full feedback.

In my first study, I compared *Room Pointing* (intrinsic and extrinsic feedback) with three types of navigation-based interactions (intrinsic, extrinsic, and system feedback). The results showed that people can use *Room Pointing* at the same level of accuracy than screen- and touch-based interactions as long as system feedback is provided (see 5.4.2, blocks 1 – 4). Without system feedback, selection accuracy dropped significantly. This result is expected to some degree as system feedback allows people to review the accuracy of their selection before confirming it. My second study, where I compared *Room Pointing* with *Ray-casting Air-pointing*, however, contrasts the finding from my first study. The results showed that participants achieved higher selection accuracy without system feedback. A $2 \times 4$ RM-ANOVA (feedback-type x block) confirmed that this increase in accuracy was significant: $F(1,11) = 5.93, p < .05$. There are several possible reasons for this discrepancy in results between the two studies, such as a higher level of proficiency in the second study due to the higher number of performed selections or a better training regime by alternating training- and trial-blocks (i.e., full system-feedback during training-blocks but only intrinsic and extrinsic feedback during trial-blocks). For a definitive explanation, a more specific study would be necessary. There are, however, several factors that did not seem to play a role, most notably the higher number of mappings in the first study (see 8.2.3 for a detailed discussion).

In summary, my results show that room-based interaction offers a broader variety of feedback channels than navigation-based interaction. With room-based interaction people can choose how much feedback they want to receive and how much attention and effort they want to dedicate to

the HEI-task. That means that people can, for example, trade off selection accuracy for system-feedback-free interaction (e.g., when there is no output medium for system feedback available). The magnitude of this trade-off, however, remains unknown. My second study showed that people can use *Room Pointing* without system-feedback while still achieving selection accuracy of above 95 %. Overall, people can use room-based interaction in a more flexible way than traditional touch-based HEI.

### *The Advantage of Real-World Proxy-Objects in Room-based Interaction*

The results of my first study, where I compared real-world proxy objects (*Room Pointing*) with screen-based proxy buttons (*Screen Pointing*) show that there is no clear advantage of using real-world proxy objects as long as system-feedback is provided. In my second study, I compared two selection techniques that do not require system feedback: *Room Pointing* and *RCAP*. The results from this study (see 6.4.1) clearly show the advantages of using real-world objects as selection proxies for digital artifacts: they are easier to remember and more accurate to point at than virtual, invisible regions in the environment. Analyzing early pointing errors (during Trials 1) showed that participants performed fewer errors with large magnitude when using real-world proxy objects (see 6.5.1). This is an indicator that people can remember associations between digital artifacts and real-world objects more quickly than between digital artifacts and virtual, invisible target zones. Analyzing late pointing errors (during Trials 5) for correct pointing gestures showed that participants performed gestures more accurately when using real-world proxy objects (see 6.5.1). This indicates that people can point more accurately toward real-world objects than virtual, invisible target zones.

The disadvantage of using real-world proxy objects is that selections generally only work in the environment that contains the real-world objects (although exceptions might be possible, see 8.3.4). By comparison, body-centric virtual proxy zones (e.g., in *RCAP*) function in every environment. While this might sound like a shortcoming of room-based interaction, I do not necessarily consider this to be a crucial disadvantage in the context of HEI. I argue that in many HEI scenarios, people actually want to change some property of the environment they are currently in, for example by turning on lights or changing the output of a screen. The interaction with the environment and the result of this interaction occur in the same spatial location, and changing the state of the environment while being outside of the environment oftentimes does

not make sense. I admit, however, that there are some scenarios in which body-centric proxy virtual zones are more useful than real-world proxy-objects.

In summary, the results from my studies show that room-based interaction offers several advantages over existing techniques for HEI, such as in-place, touch-based, and navigation-based interactions. In contrast to touch-based techniques, room-based interaction offers device-free interaction, which can be useful when people do not want to touch or hold an interaction device. In contrast to navigation-based interaction, room-based interaction allows for system-feedback-free interaction, which can be useful when people do not want to shift their attention to the supporting HEI-task or when the environment does not provide channels for outputting system feedback. Room-based interaction offers people more choice for adapting HEI to their particular needs and requirements than aforementioned existing HEI techniques. By doing so, room-based interaction solves the problem that existing HEI techniques do not offer device-free interaction.

### 8.1.3 Public Visibility of Room-based Interaction

In the introduction, I argued that room-based interaction solves the problem that current techniques for HEI hide interactions from other people in the same environment. Many interactions in a smart environment change some property of that environment and thus affect everyone located within. Generally, co-located people should know about these changes, because of collaboration efficiency or simple courtesy.

In my third study (see 7.4.1), I compared observability and identifiabilty of differently sized gestures. I hereby focused on gesture sizes typical for touch-based interaction and gestures sizes typical for room-based interaction. My results show that full-arm pointing gestures are more visible than phone- or tablet-sized gestures. While this is not surprising, numerous additional conclusions that benefit interaction designers can be drawn from this study. First, my results establish a base-level for gesture visibility: even small smart-phone- and tablet-sized gestures are quite observable and identifiable (see 7.4.3). Second, my results indicate that other properties of a gesture besides size, such as gesture morphology, can have a profound influence on gesture visibility (see 7.5.2). Third, my results indicate that gesture size does not have to be considered as an ordinal property (e.g., small and large) when designing for publicity but instead can be seen as a scalar property that allows interaction designers to fine-tune the amount of publicity they want to achieve (see 7.5.1 and 7.5.2). Interaction designers can use the results from this study to

estimate the visibility (i.e. privacy or publicity) of their interaction technique or they can use the results to design gestures for their interaction technique to achieve the desired level of privacy or publicity. Designers also have to be aware that there is a strong correlation between gestures size and gestures publicity: it could be difficult to design large private or small public gestures (see 7.5.1).

More generally, full-arm gestures have the potential for creating higher levels of awareness through consequential communication than phone- or tablet-sized gestures. I looked at two out of the three steps for awareness creation (see 2.2.4): perception of action (or gesture observability) and comprehension of the situation (or gesture identifiability). For these two steps, my results showed that (large) mid-air full-arm pointing gestures have a higher probability of being observed and correctly identified by other people in the environment than phone- or tablet-sized gestures. While this does not necessarily mean larger gestures create higher awareness, it seems plausible that they will.

In the introduction, I argued that room-based interaction has the potential of bringing publicity of interactions back to HEI, which had been present with in-place interfaces but was lost with navigation-based interaction (see 1.1.3). My third study showed that the mid-air full-arm pointing gestures used in room-based interaction are more visible that smaller gestures used, for example, in touch-based interaction, and are, thus, likely to produce higher levels of awareness for other co-located people in smart environments. Room-based interaction is a viable solution for creating public HEI techniques.

## 8.2 Summary of Secondary Findings

In addition to the three main advantages for room-based interaction that I argued for in the introduction, several findings emerged from the design of each of the three studies in my dissertation. In this section, I go through three of the most important ones.

### 8.2.1 Mental Model of Pointing-based Interaction

In my second study (Chapter 6), I compared *Room Pointing* with *RCAP*. The results showed that people's mental model about different interaction techniques might differ substantially from what the interaction designer anticipated (see 6.5.2). When analyzing mid-air full-arm pointing gestures (see 3.2.2 and 3.2.3), I was convinced that room-based interaction with its real-world

proxy-objects should lead to a different mental model in people than interaction techniques using body-relative proxy-zones (e.g., *RCAP*, *Virtual Shelves*). The post-technique questionnaires in my second study (see 6.5.2) showed, however, that some participants did not conceptualize virtual, invisible target zones for *RCAP* but instead used real-world objects within the zone as pointing targets and memory aids. While priming plays undoubtedly a role in shifting some participants' mental model, other factors might be important, too. One factor might be that participants made a conscious decision to use real-world proxy objects in *RCAP* because they felt that the association between digital artifact and real-world object was easier to memorize than between digital artifact and virtual, invisible proxy-zone. Another factor might be that participants felt that pointing at a virtual, invisible proxy-zone would be less accurate than pointing a real-world object; P3 corroborated this by saying that "eventually, I was able to fine-tune my accuracy and stopped paying attention to objects".

Overall, I not only showed that the mental model underlying room-based interaction helps people to better learn associations between digital artifacts and selection proxies but I also argue that associations between digital artifact and real-world proxy-object are more natural and intuitive than associations between digital artifact and virtual, invisible proxy-zones.

### 8.2.2 Structure of the Storage Space

The first study (Chapter 5), I compared hierarchical (*Touch Scroll*) and flat (*Touch Flat*, *Screen Pointing*, *Room Pointing*) storage spaces. The results showed that the structure of the storage space significantly influences people's performance with a selection technique. The results suggest that interaction techniques with a flat storage space have a selection speed advantage over techniques with a hierarchical input space, likely because a flat storage space does not require users to spent time on navigation. Therefore, I would recommend that interaction designers who focus on achieving fast selection speed should try to keep the storage space of their interaction technique flat. Naturally, the storage space on any interaction technique is finite, so there is a limit to the number of digital artifacts that can fit in the input space. With decreasing size, each element in the storage space becomes more difficult to select accurately, which ultimately either leads to reduced selection accuracy or selection speed. This so-called speed–accuracy-tradeoff is an inherent attribute in the human motor system (Schmidt, Zelaznik, Hawkins, Frank, and Quinn, 1979) and, thus, frequently observed in HCI-research (MacKenzie,

1992). In my study, however, the number of elements was too low to show this effect (see 5.4.1 and 5.4.2) despite buttons during the trials$^{+}$-condition being slightly smaller ($1.0\ cm$ wide) than on most current smart phones (e.g., Nexus 4: $1.2\ cm$ wide).

This result demonstrates one of the potential advantages of room-based interaction: its use of the environment as large and flat input space. The flat nature of the input space guarantees that people are able to make fast selections, the large size of the input space guarantees that people are able to store a large amount of selection proxies before the speed–accuracy trade-off has a significant influence on people's selection speed or accuracy. For a more detailed discussion on the maximum amount of proxies in an environment, see 8.3.1.

### 8.2.3 Accuracy of Room-based Interaction

The first study (Chapter 5), I compared room-based with navigation-based interaction. The results showed that selection accuracy with room-based interaction is lower (88 %) than with other techniques (e.g., *Screen Pointing*: 98 %). The second study (Chapter 6), where I compared *Room Pointing* with *RCAP*, confirmed the data from the first study, although selection accuracy was slightly higher (92 %). This difference in selection accuracy for *Room Pointing* between the first and the second study might be due to less training in the first study, the difference in training regime, or some other artifact in the experiment design. I will not get into more detail as it is of less interest to the overall discussion of accuracy of room-based interaction. The main question is what properties of room-based interaction determine people's accuracy.

When looking at the distribution of proxy objects (see Figure 100), two possible explanations could be that either the size of the target area around the proxy object or the density of proxy objects in the region could influence selection accuracy. I calculated target size as the percentage of the imaginary sphere around the person that a target is covering (see 4.2.5) and target density as the average angular distance between a target and its direct neighbors.

#### *Target Size*

The rationale for using target area as a predictor for selection accuracy is that larger target areas might be easier to point at. Target sizes in the first study ranged from 0.3 % to 14.4 % ($\mu =$ 3.3 %) and from 1.1 % to 5.2 % ($\mu = 3.3$ %) in the second study.

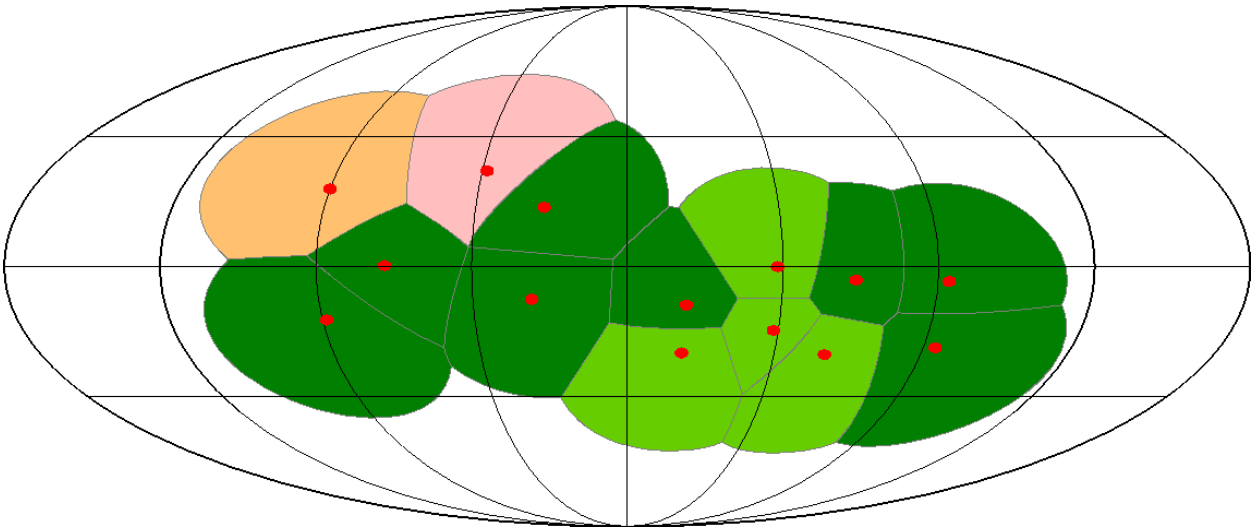


**Figure 100: Selection accuracy per real-world proxy object; dark green: 95 – 100 %, green 90 – 95 %, orange: 80 – 85 %, and red: 75 – 80 % accuracy**

A linear regression analysis showed little correlation between target size and selection accuracy: $F(1,42) = 0.1, p > .7$, with $R^2 = .00$.

### *Target Density*

The rationale for using target density is that it might be more difficult to point at a target that is close to other targets. Target density ranged from 22° to 72° ($\mu = 40°$) and from 24° to 45° ($\mu = 34°$) in the second study.

As for target area, a linear regression analysis showed little correlation between target density and selection accuracy: $F(1,42) = 0.3, p > .5$, with $R^2 = .01$. Figure 101 shows the relationship between target size and selection accuracy for both Study 1 and Study 2.

In the second study, the target area with the lowers selection accuracy could be considered a $2\sigma$-outlier. Removing it from the data, however, does not change the fact that statistical there is little correlation between the target size, target density, and selection accuracy).

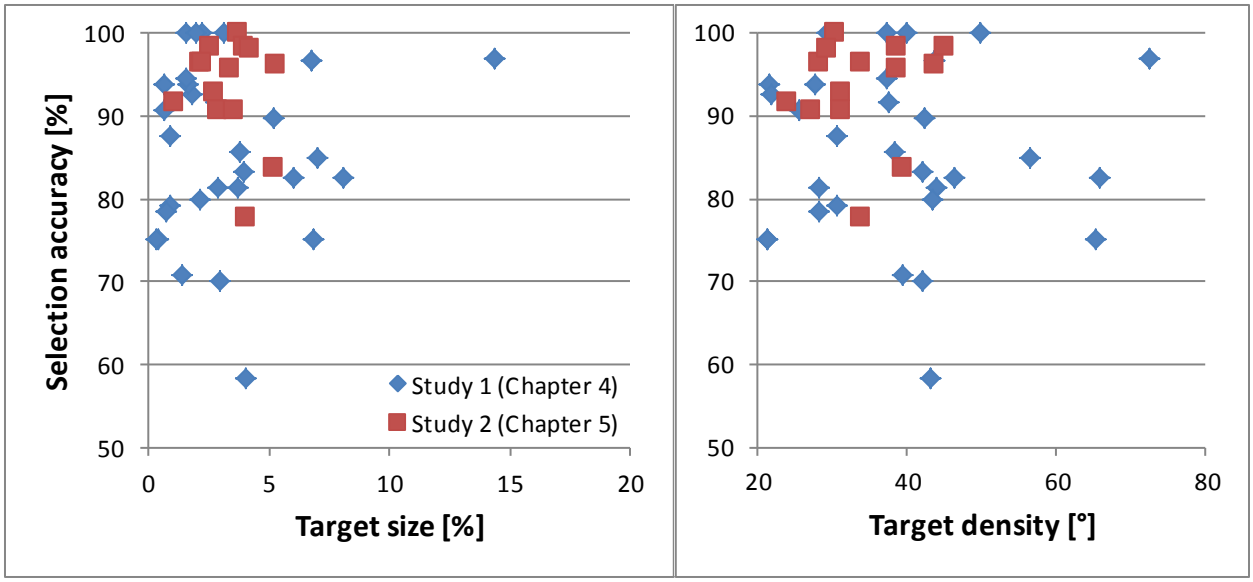



**Figure 101: Selection accuracy as a function of target size (left) and target density (right)**

Although, this lack of correlation might seem surprising, it is in line with previous research, which showed that people's distal pointing errors are below 2.5° when the pointing target is in foveal vision (see 2.4.4). The size of the target areas in both studies was simply too large for having an effect due to limitations in people's pointing capabilities. This finding is important because it shows that people's pointing capabilities are not the limiting factor, even when having 40 or potentially more proxy objects in the environment. Figure 102 illustrates this point by showing people's average pointing errors (red) overlaid on top of the target zones. This figure shows that people rather adapt their pointing strategy depending on target area and density: for small target zones in crowded regions, people increase their pointing accuracy, while they are less careful when pointing toward large target zones. A linear regression analysis between target size, target density, and people's pointing errors show that target size ($F(1,42) = 74.2, p < .001$, with $R^2 = .64$) and target size x target density ($F(2,41) = 295.5, p < .001$, with $R^2 = .673$) are good predictors for people pointing effort, while target density is less predictive ($F(1,42) = 16.7, p < .001$, with $R^2 = .28$).

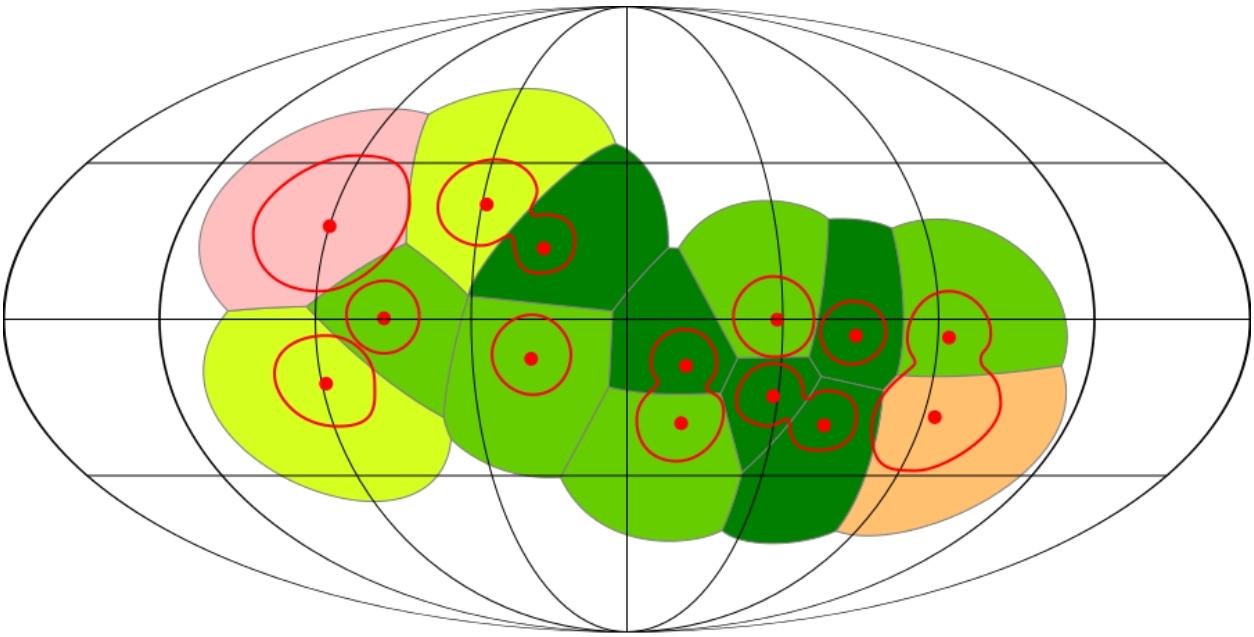


**Figure 102: Pointing error per real-world proxy object for Study 2 / Trials 1 – Trials 5; dark green: $5° - 10°$, green $10° - 12.5°$, yellow: $12.5° - 15°$, orange: $15° - 17.5°$, and red: $17.5° - 20°$ error; $\mu = 11.9°$; red circles show average pointing errors per proxy object**

In conclusion, I argue that my first two studies did not push against the limits of people's pointing skills and that the number of mappings in the environment could be further increased without significantly affecting people's selection accuracy. Pointing errors seem to be caused by people willingly adjusting their pointing accuracy depending on the target size and not by people's motor skill limitations. With sufficient effort and training, there is little reason why people's selection accuracy should be lower with room-based than with navigation-based interaction.

## 8.3 Additional Findings and Discussions

Besides the main findings presented above, there are several additional findings that emerged from the results of the three studies of my dissertation.

### 8.3.1 Limitations for the Number of Proxy Items in Room-based Interaction

There are three factors that limit the number of digital artifacts that people can store with room-based interaction: the number of real-world objects in the environment that can serve as proxy

objects, people's ability to remember associations between digital artifact and proxy object, and people's pointing accuracy.

### *Limitations due to pointing accuracy*

Previous research has shown that people's pointing error is below 2.5° for targets within foveal vision (see 2.4.4). The question is now how many (circular) pointing targets $N$ with radius $r = 2.5°$ would fit on a sphere. This seemingly simple problem is currently unsolved, although it has received attention in biology, where it is known as *Tammes problem*, and in physics, where it is known as *Thomson problem* (Aste and Weaire, 2008). In these two fields the problem is usually reversed as "what is the maximum (angular) distance $r$ between $N$ points on a sphere". While this problem has been solved for some values of $N$, it has not been generally solved. There is, however, a function for calculating an upper limit of the maximum (angular) distance (Tóth, 1943):

$$r \leq \sqrt{4 - \csc^2\left(\frac{\pi N}{6(N-2)}\right)}, N \epsilon \mathbb{N}, N \geq 3 \quad (1)$$

This function can be solved for $N$ to calculate the number of (circular) pointing targets for a given target radius $r$:

$$N = \left\lfloor \frac{12 \csc^{-1}\sqrt{4 - r^2}}{6 \csc^{-1}\sqrt{4 - r^2} - \pi} \right\rfloor, r \epsilon \mathbb{R}^+, r \leq \sqrt{3} \quad (2)$$

For $r = 2.5°$, $N \approx 7600$, which means that with optimal pointing performance people should hypothetically be able to accurately point at 7,600 proxy targets. More realistically, my second study showed that people reached 95 % selection accuracy with *Room Pointing* during the last trial-phase with 14 proxy targets of no more than 34° in radius (see 6.4.1). In fact, the mean distance between all adjacent targets (i.e., targets that shared one border) was 32.5°, which is equivalent to an average radius of 16.25° per target (see 6.3.2). Given these conditions ($r = 16.25°$ and tolerated selection error of 5 %), the number of pointing targets $N \approx 180$, which is still much higher than the number of selection proxies on a touch-based interfaces.

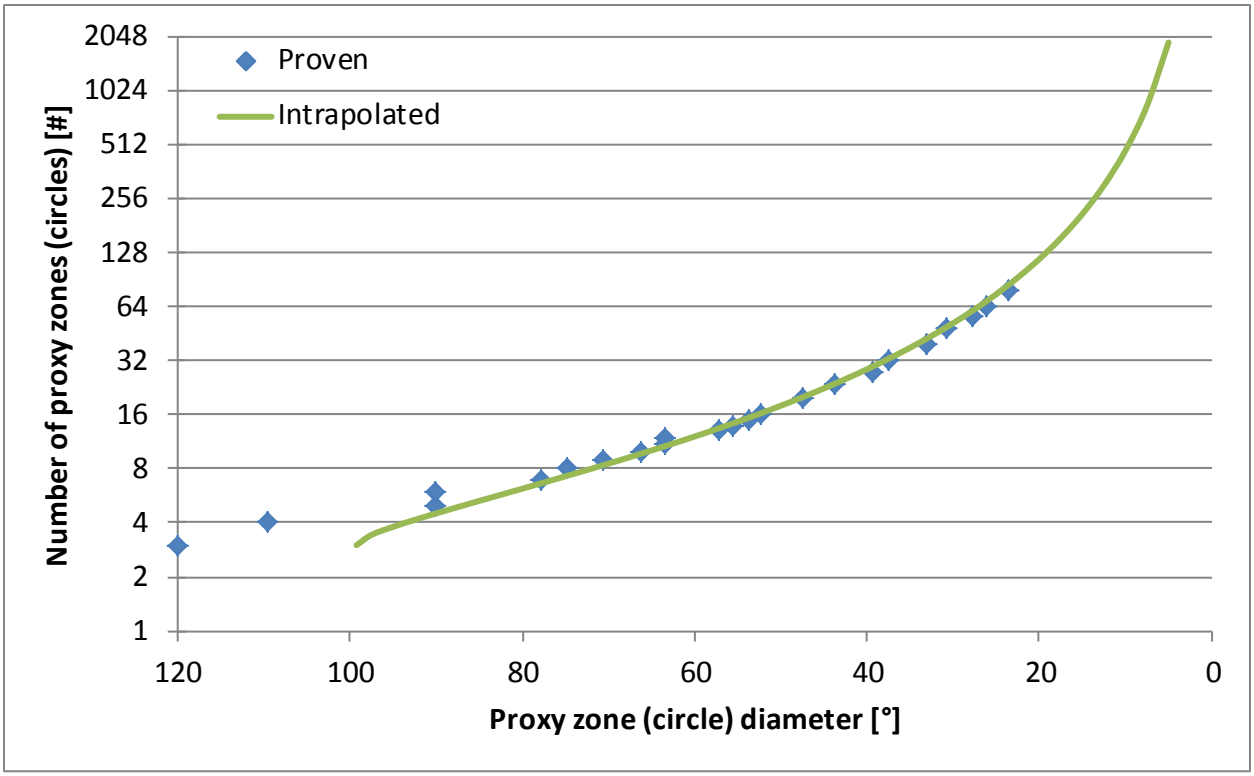



**Figure 103: Potential number of proxy zones for a given zone diameter; y-axis logarithmically scaled (values adapted from Erber and Hockney, 2007, pp. 556–573)**

### *Limitations due to Association Recall*

Assessing how well people can remember associations between digital artifact (cue) and real-world proxy-object (response) is difficult as no research for this particular scenario exists. There are, however, some standardized tests involving associative memory for other scenarios. Semantic fluency, for example, is the number of words of a cued category that a person can list within a short time, usually 60 seconds; the semantic fluency of healthy adults is around 20 (Troyer, Moscovitch, and Winocur, 1997).

In my first and second study, participants only had to memorize 7 and 14 mappings between digital artifacts and real-world proxy objects. In an early study, however, I asked 9 participants to memorize 30 mappings that were created by the participants themselves. Participants achieved a selection accuracy of 88 % ($\sigma = 9.0$ %), which shows that they were able to make selections on a similar level of accuracy than in my first study (see 5.4.2). The results from this experiment,

however, are not directly comparable to my first and seconds study in this dissertation since the study system itself was quite different in terms of tracking hardware (see 4.1) and software (see 4.2.4). Having that said, the results still demonstrate that people can learn a large amount of associative information within a short timeframe and remember it accurately.

### *Limitations due to the Lack of Real-world Proxy Objects*

Arguably, every domestic environment is populated with hundreds of real-world objects. However, not all of these objects can be used as proxy objects because, as mentioned previously, real-world proxy objects should be unique and static (see 3.2.2). The reason for them to preferably be static is that otherwise people might have difficulties finding the proxy object in the environment (e.g., a key-chain might not be ideal, although the location where the key chain usually lives might be), and they should be unique so that they are not easily confused with other objects (e.g., a particular brick might not be ideal, although the brick wall in its entirety might be). In addition to these cognitive factors, potential proxy-objects should also not occlude each other as this makes pointing gestures ambiguous. This ambiguity, however, can easily be avoided by selecting proxy objects that are close to the wall since people mostly stay near to the center of a room. Admittedly, there is little research about the number of potential proxy objects in people's environments, especially when considering the narrow definition and requirements for proxy objects in the context of my dissertation. Most existing literature is either related to the number of objects that people suffering from dementia interact with (e.g., Galasko, Bennett, Sano, Ernesto, Thomas, Grundman, and Ferris, 1997, as part of health care research) or the number and type of objects children play with (e.g., Rheingold and Cook, 1975, as part of gender studies). In the absence of other research, I present two case studies that demonstrate the high number of potential proxy objects. The first example is a 360° view of a domestic environment—the author's living room. I marked 50 potential proxy objects that are all satisfy the three above mentioned requirements: they are static, they are unique, and they show little occlusion from the locations where people would normally stay. In addition, all of these real-world objects have rich semantic meaning to the author.

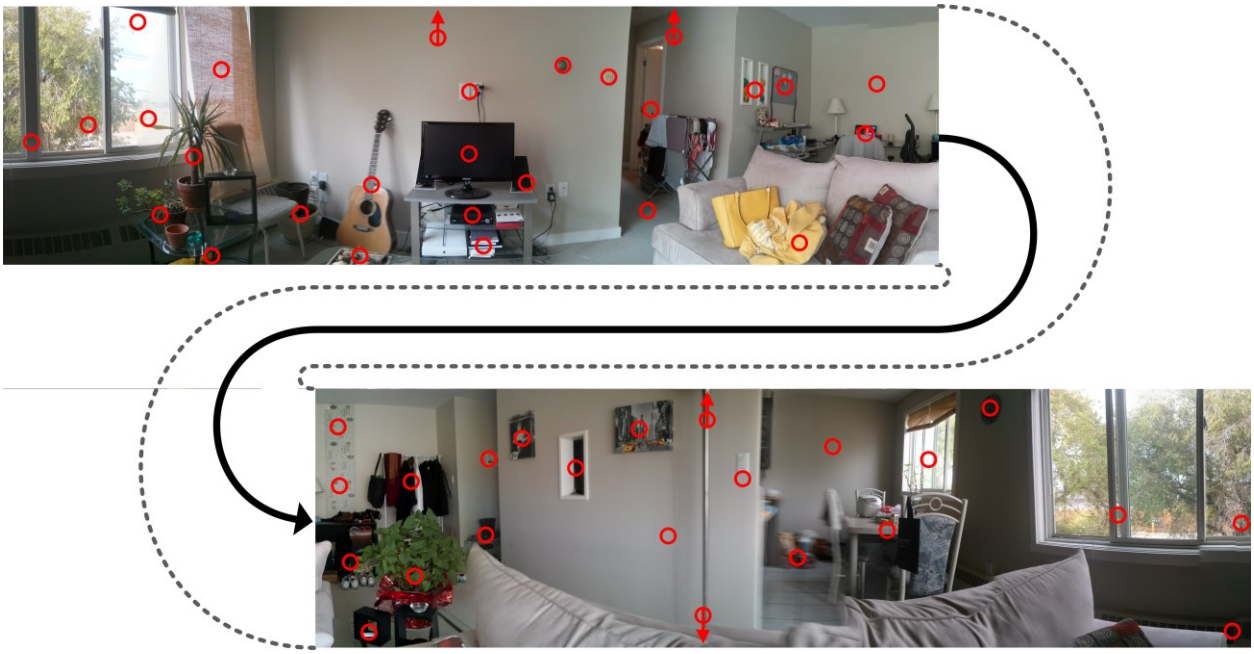


**Figure 104: Example of 50 potential real-world proxy objects in a domestic environment[6]**

The second example is a 360° view of an office environment—the lab in which all of the user studies took place. Again, I marked 50 potential proxy objects that are all satisfy the three above mentioned requirements. Like before, all of these real-world objects have rich semantic meaning to the author.

Overall, I argue that people will not have problems finding hundred or more real-world objects with rich semantic meaning in an environment that they are familiar with.

[6] From left to right: living room window, Spicy Garden restaurant, sky, Safeway, Nora's plant, Yucca cane, small glass table, broken curtain, empty pot, large table, guitar, left light, main power outlet, TV, cable box, books, speaker, thermostat, fire alarm mute, hallway, bathroom, fire alarm, footballs, pin board, loveseat, computer, space over computer, wall poster (top), wall poster (bottom), shoe cartons, where the remote control normally is, coat hanger, poinsettia, show rack, where the keys normally is, Marilyn Monroe, kitchen, Time Square, stove, lamp foot, lamp head, light switches, rice bags, empty wall in kitchen, kitchen table, kitchen window, clock, Starbucks, broken tree, radiator

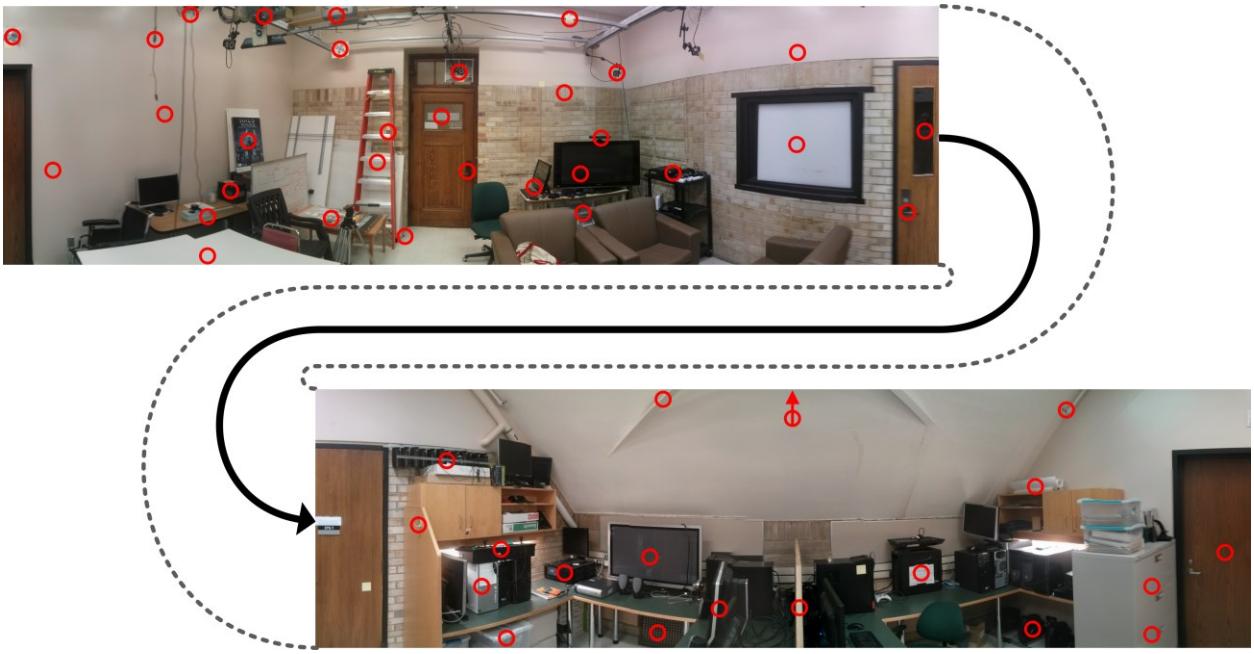


**Figure 105: Example of 50 potential real-world proxy objects in a lab environment[7]**

### 8.3.2 Selection of Real-world Proxy Objects

One aspect that sets room-based interaction apart from most existing selection techniques is that people have to create their own set of associative mappings between digital artifacts and real-world objects. Existing research suggests that these mappings would highly depend on the amount of meaning that each individual can construct between artifact (cue) and proxy-object (response) (see 2.5.5).

In an early experiment, I gave 9 participants a list of 30 TV shows and asked to create mappings between these shows and real-world proxy objects in the lab environment. Out of the 270 recorded mappings, 221 were unique, and only 49 mappings appeared more than once. The most common mapping was between a show called "The Hour" and a wall clock; 8 out of 9

[7] From left to right: picture over office door, space right of the office door, dangling cable, big empty wall, main network switch, brown desk, projection table, Polhemus system unit, Halo-poster, rear-projector, small coffee table, front-projector, scissor drawing, red ladder, thermostat, door stop, front-door window, New York, front-door handle, TV-PC, space above main TV, yellow sticky tag, main TV, media bin, Kinect sensor, corner camera, XBox 360, two-way mirror, space above two-way mirror, lab-door handle, lab-door window coat hanger, speaker array, white computer, Street Figher controls, old keyboards, server, rear TV, scrawl-space access, crawl-space access, cross-beams, eye-tracker, ceiling window, separator wall, Gregor's PC, bike pedals, rolled-up posters, sprinkler, cabinet (bottom), cabinet (top), office door

participants were using the clock as real-world object for one of their mappings, and 6 out of 8 assigned it to that show. Overall, participants used more than 100 different real-world objects for 270 mappings. Only 43 real-world objects were used by more than one participant. These 43 real-world objects contributed for 185 out of the total number of 270 mappings. Only one real-world object—an office door—was used by all nine participants. It was, however, mapped to seven different TV shows. These results were in line with existing research (see 2.5.5): most mappings should differ between individuals since everyone used their personal experience to create meaning; a few mappings were expected be more common since there is a body of shared experience and, thus, meaning that every person was tapping into (a clock tells time ↔ an hour is a measurement of time ↔ "The Hour").

Overall, the results from this experiment show that people are able to create meaningful associations between digital artifacts and real-world proxy objects. The degree of variety, creativeness, and individualism when creating these associations is remarkable. This also has some ramifications on awareness creation with room-based interaction. As I described above, three steps are necessary to create awareness: perception of an action, comprehension of the situation, and projection of the future status (see 2.2.4). As I showed in my third study, room-based interaction support people well in the first two steps. To complete the last step, projection of future status, people have to understand what the gesture means, i.e. which digital artifact is associated with each real-world proxy-object. The results from this early experiment indicate that people generally do not have much shared meaning (see 2.5.5), which would make the projection of the future status difficult. This means that people would have to learn each other's mappings, a task that could be tedious but is certainly not impossible: in both study 1 and 2 participants successfully learnt the associations that I provided them. To draw a final conclusion about learnability of others' associations, future research is necessary.

### 8.3.3 Designing and Deploying Room-based Interaction Techniques

This section summarizes the design consideration that are spread out through my dissertation. The purpose of this section is to be a guide for interaction designers that seek to implement a room-based interaction technique.

The association between digital artifact and real-world proxy object is one of the two core components of room-based interaction (the other one is the use of mid-air full-arm pointing gestures). In order to support memorization of these associations, an environment has to contain real-world proxy objects that are rich in semantic meaning to the user. In addition, proxy objects have to be distinguishable from each other, i.e. carry different meanings, so that users do not confuse them easily. Designers should, however, not underestimate people's abilities to concoct semantic connections between seemingly unrelated digital artifacts and real-world objects.

In order to support pointing gestures, proxy objects should be static within the environment, i.e. not move around. If they are static, people can use their spatial memory for finding the object quickly. This implies that room-based interaction works best in familiar environments, i.e. environments in which people have spent enough time to build up sufficient spatial memory. Last, proxy objects should be spaced out as much as possible to allow for some inaccuracy in people's pointing gestures.

When picking proxy objects, the interaction designer or user should consider occlusion, which occurs when one proxy objects lies between the user and another proxy object. This problem can quite easily arise as people move around in the environment. The best way to avoid occlusion is only using real-world objects as proxy that are close to the edges of the environment, i.e. the walls, ceiling, and floor. When searching for potential occlusions, it might be worth considering from which locations in the environment people are most likely to perform HEI. Just like with deictic pointing, people are, however, able to point around an occluding proxy object.

### 8.3.4 Limitations, Generalizability, and Application Areas

In all of my three studies, I made several assumptions that limit the generalizability of my statements regarding room-based interaction.

#### *Lab-based Studies versus Real-world Deployment*

I never deployed *Room Pointing* in a real domestic or office environment. It is common practice in HCI-research to use lab studies for many reasons. First, lab studies are easier to conduct because lab study system do not require the stability and sophistication of a field study system, which add complexity to the system without benefitting the HCI-aspect of the research. Second, lab studies are more controlled, which means they reduce the number of confounding factors that

contribute to the measured outcome of the study. Third, lab studies are faster to set-up and conduct, so they generate more data, which in turn makes the results more generalizable. The scope of application for lab and field studies varies: lab studies are usually used for collecting (quantitative) performance data, whereas field studies are used to (qualitatively) evaluate use cases and behavior. My research focused on the qualitative performance aspect of room-based interaction, so lab-based studies were the logical choice. Deploying *Room Pointing* in the field could be an interesting option for future research. Anecdotally, I feel confident that people would use room-based interaction in their homes as many participants expressed their enjoyment of using *Room Pointing*.

Another common problem with lab-based studies conducted at universities is that participants are not representative of the general population since they are usually younger and more technically versed than an average person. I believe, however, that this bias does not affect the generalizability of my results: the cognitive , spatial, and motor tasks that I required participants to perform were basic enough so that any healthy adult should be able to perform them, and participants generally had little to no previous experience with advanced gesture-based interfaces, such as room-based interaction or *Ray-casting Air-pointing*.

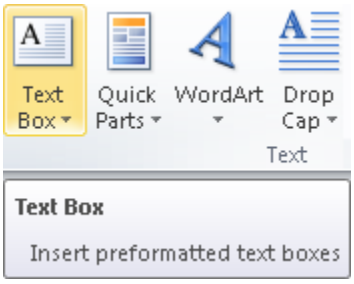
### *Complexity of Selection Task*

In my dissertation, I decided to focus on single-selection tasks, i.e. task that can be completed with a single artifact selection from a larger group (3.3.1). While this decision omits a majority of possible and more complicated interactions, it does not reduce the contribution of my work. This work is the first comprehensive user study on room-based interaction and thus follows the common practice in HCI-research to use basic and fundamental tasks for the initial investigation of interaction techniques.

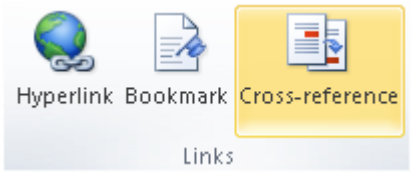
It is, however, possible to speculate about using room-based interaction for more complex selection tasks. Examples for such tasks could be selecting one continuous value (e.g., speaker volume), selecting more than one item (e.g., dialing a phone number), or selecting from a larger subset of digital artifacts (e.g., “TV stations” → “CNN”). While the last example goes against the design recommendation for room-based interaction to retain a flat input space for as long as possible (8.2.2), people might still be interested in using room-based interaction in such way. *Charade* showed a possible solution for using multi-stage interaction in room-based interaction

(2.2.3). The sign-based gestures used in *Charade*'s second stage could be easily adopted to be used with room-based interaction; for example, one could use a pointing gesture toward the ceiling lights to turn them on and then a pinch-like gesture for dimming the lights. Another possibility could be to use chorded sign-based gestures similar to the ones in *Marking Menus* (Kurtenbach, Sellen, and Buxton, 1993), which would allow for a sequence of selections similar to browsing through sub-menus.

### *Domestic-, Office-, and Other Scenarios*




Initially, I envisioned room-based interaction in the context of domestic and office settings. The three scenarios described in 3.3.2 outline typical domestic scenarios. In these settings, I envision room-based interaction as a preferred choice for HEI during non-computer-based primary tasks. Room-based interaction could be particularly helpful as it helps people to continue working on their primary task with minimal interference from the supporting HEI-based task.




In office settings, where people's primary task is mostly already computer-based, room-based interaction can still be helpful. People could use it, for example, as an accelerator mechanism that is faster to execute than a mouse click on an icon and easier to remember as a keyboard shortcut. Keyboard shortcuts might still be the preferred choice for constantly performed actions, such as copy and paste (CTRL-C, CTRL-V) and icons the preferred choice for rarely executed commands, such as INSERT TEXT BOX. The number of possible keyboard shortcuts, however, is limited, and many keyboard shortcuts are difficult to memorize (ALT-SHIFT-X to mark entry). In office settings, room-based interaction could fill this gap by providing an additional input space for easy to remember command accelerators. (While writing this dissertation, I dearly missed having a shortcut for INSERT CROSS-REFERENCE.)

Similarly to the domestic scenarios, there are other work environments in which switching between primary and supporting task is time-consuming or even impossible. In operating theaters, for example, surgeons are not allowed to use their hands to control computer systems as this could negatively influence the sterility of their gloves, increase the chance of post-surgical infections, and ultimately lead to sepsis and death in patients (O'Hara, Gonzalez, Penney, Sellen,

Corish, Mentis, Varnavas, Criminisi, Rouncefield, Dastur, and Carrell, 2014). In this particular scenario, the device-free nature of room-based interaction would allow surgeons to control computer systems directly, thus reducing the chance of miscommunication within the surgical team while creating group and situation awareness through the use of large mid-air full-arm pointing gestures

### *Room-based Interaction in Arbitrary, Unknown, or Non-static Environments*

In my recommendations on the design of room-based interaction, I argued for using static real-world proxy objects (3.2.2), which are numerous in most domestic or office environments (see 8.3.1). Given people's generally excellent spatial memory (see 2.5.3), however, one could imagine using room-based interaction outside of the familiar environment where the mappings were originally created, for example, their living room. This would allow people to use room-based interaction in any environment. Whenever people would want to issue a system command, they would simple imagine themselves being at a certain spot in said familiar environment, for example, standing in front of the couch, facing the TV. People would then simply perform a pointing gesture toward the real-world proxy object, for example, the living room door, solely guided by spatial memory.

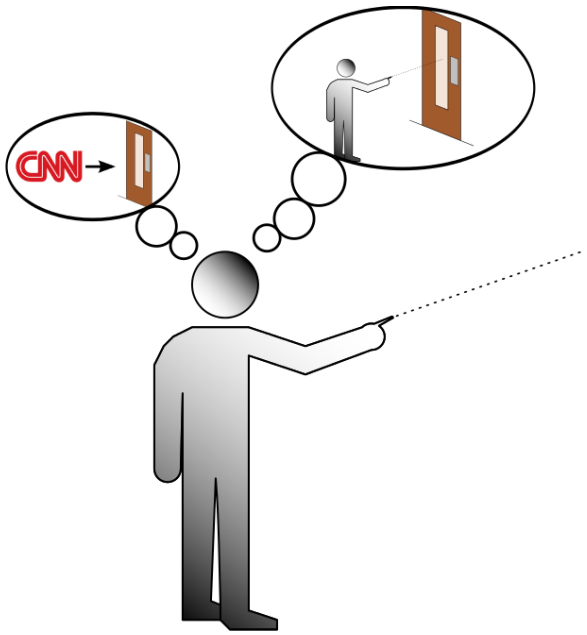



**Figure 106: Performing *Room Pointing* without being in the environment**

In this use case, *Room Pointing* could be considered similar to *Ray-casting Air-pointing* in that people would point toward proxy zones. While these zones remain invisible and virtual for *RCAP*, I would argue that they are not virtual but rather imaginary for *Room Pointing*: they are still represented by a real-world proxy object and, thus, all the advantages of having real-world proxy objects remain (see 8.1.2), the real-world object just happened to be no there and has instead to be imagined.

I find this concept intriguing as it would solve one of the major problems of RCAP: memorizing the association between digital artifact and selection proxy. Given the discussion on feedback-channels (see 2.4.2 and 8.1.2), I would expect selection accuracy to be lower than in the traditional *in situ* version of room-based interaction. *Ex situ* room-based interaction would be an interesting topic for future research.

# Chapter 9 Conclusion

Controlling digital artifacts in smart environments is an increasingly frequent task as these environments are becoming more common. The history of, for example, smart phones and tablets has shown that oftentimes it is not the functionality and capabilities of a technology that determine its success, but the quality of the interaction with the technology. For Human-Environment Interaction this means that it must integrate itself into people's life so that it supports their daily routines instead of interfering with them. Providing this kind of seamless interaction might be a key factor that will decide the success of smart environments. Unfortunately, neither HEI through in-place interaction, such as wall-mounted buttons nor through navigation-based interfaces, which are frequently used on smart phones and tablets, integrate themselves seamlessly into many of people's daily routines.

In this dissertation, I presented **Room-based Interaction** as an alternative to using in-place interaction or navigation-based interfaces for Human-Environment Interaction. With room-based interaction, people can use mid-air full-arm pointing gestures toward real-world proxy-objects to interact with smart environments. Based on an in-depth review of existing research, I created a conceptual framework for analyzing different types of human-environment interactions. This framework helps understanding the cognitive processes involved in producing pointing-based selections. I then designed, implemented, and evaluated multiple prototypes of room-based interaction, which led to the creation of *Room Pointing*, a novel room-based interaction technique for making selections in smart environments. The design of room-based interaction was informed by my conceptual framework. Then, I conducted three user studies where I evaluated *Room Pointing* and other existing pointing-based and touch-based selection techniques. The main goal of these studies was to provide evidence that supports my assumptions about room-based interaction and verify my conceptual framework. The focus of

the first study was comparing existing navigation-based interaction with room-based interaction and assessing the influence of differences in storage space, selection mechanisms, and proxy types on people's selection performance. The second study focused on comparing two pointing-based interaction techniques and assessing the influence of proxy types on people's selection performance and learning rate. In the third study, I investigated the opportunity for using room-based interaction to increase awareness between co-located people. Finally, I discussed how my findings confirmed my initial assumptions about room-based interaction, how they verified my conceptual framework, and what additional conclusions can be drawn to the design and application of room-based interaction.

## 9.1 Contributions

There are four main contributions of this dissertation.

First, my dissertation establishes the usefulness of real-world objects as selection proxies in smart environments, which has not been rigorously investigated yet in existing literature. The use of real-world proxy objects in room-based interaction demonstrates two of the advantages. It grants people the ability to make selections more accurate than with other selection techniques that are also using mid-air full-arm pointing gestures (e.g., Ray-casting Air-pointing). In addition, it helps people to learn associations between digital artifacts and proxies faster than with selection techniques using virtual, invisible proxy zones (e.g., Virtual Shelves and RCAP).

Second, my dissertation shows the usefulness of mid-air full-arm pointing gestures as interaction mechanisms in smart environments. Room-based interaction proofs that people can make selections as fast and as accurate as with touch-based interfaces. It also demonstrates that people can achieve high performance when using room-based interaction system-feedback-free. Finally, it suggests that that people can use room-based interaction eyes-free as well.

Third, my dissertation presents a conceptual framework for assessing pointing-based interaction techniques. This framework accurately predicted the results in my user studies; the results of the user studies thus validated the conceptual framework. My dissertation also presents an implementation of room-based interaction called *Room Pointing*, which showcases the feasibility of room-based interaction given current hard- and software systems.

Fourth, my dissertation demonstrates the relationship between gestures size and workplace awareness through consequential communication. This information is useful for interaction designers who want to control the privacy or publicity of their interaction technique.

## 9.2 Future Work

There are several directions in which one might further investigate room-based interaction.

One would be to focus the mappings between real-world proxy objects and digital artifacts: how many associations can people remember?; how long does it take people to learn their own or other people's associations?; how high is the retention rate after days, weeks, or months? This research could extend an initial longitudinal study regarding retention recently conducted (Perrault, Lecolinet, Bourse, Zhao, and Guiard, 2015).

Another research direction would be focusing on the gesture-aspect of room-based interaction and measure people's performance with a significantly increased number of mappings, for example, more than 50, 100, or 150.

One interesting research goal would be to demonstrate the usefulness of room-based interaction in a real-life setting, i.e. deploy *Room Pointing* in a smart environment, let people it use over an extended period of time, and gather quantitative and qualitative data about people's performance, usage patterns, subjective opinions, *et cetera*.

One final area for future work would be to fully investigate the usefulness of room-based interaction for creating awareness between co-located users. This dissertation laid the foundation for this research by measuring gestures observability and identifiability, so assessing gesture interpretability would be a logical extension of my research. Future research could take two approaches to this problem: reducing gesture interpretability, which would be important for maintaining privacy or in competitive settings (e.g., video games), or increasing gesture interpretability, which would be crucial in co-located collaborative settings.

# Chapter 10 References


Gregory D. Abowd. 2012. What next, Ubicomp? Celebrating an intellectual disappearing act. In *Proceedings of the 14th conference on Ubiquitous Computing – UbiComp '12*. ACM Press, New York, NY, USA, 31–40. http://doi.org/10.1145/2370216.2370222

Jack A. Adams. 1971. A closed-loop theory of motor learning. *Journal of Motor Behavior* 3, 2, 111–150. http://doi.org/10.1080/00222895.1971.10734898

Arvin Agah. 2000. Human interactions with intelligent systems: research taxonomy. *Computers & Electrical Engineering* 27, 1, 71–107. http://doi.org/10.1016/S0045-7906(00)00009-4

Gary L. Allen. 2003. Human spatial memory: Remembering where. Psychology Press, New York, NY, USA.

John R. Anderson and Gordon H. Bower. 1980. Human associative memory: a brief edition. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Jussi Ängeslevä, Sile O'Modhrain, Ian Oakley, and Stephen Hughes. 2003. Body mnemonics. In *Workshops of the 5th conference on Human-Computer Interaction with Mobile Devices and Services – MobileHCI '03*. ACM Press, New York, NY, USA.

Perry A. Appino, J. Bryan Lewis, Lawrence Koved, Daniel T. Ling, David A. Rabenhorst, and Christopher F. Codella. 1992. An architecture for virtual worlds. *Presence: Teleoperators and Virtual Environments* 1, 1, 1–17.

Aristotle. 1973. De sensu and De memoria. Arno Press, New York, NY, USA.

Ronald T. Azuma. 1997. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments* 6, 4, 355–385.

Alan Baddeley. 1998. Human memory: Theory and practice. Allyn & Bacon, Boston, MA, USA.

Alan Baddeley. 2000. The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences* 4, 11, 417–423. http://doi.org/10.1016/S1364-6613(00)01538-2

Rafael Ballagas, Jan Borchers, Michael Rohs, and Jennifer G. Sheridan. 2006. The smart phone: a ubiquitous input device. *Pervasive Computing* 5, 1, 70–77. http://doi.org/10.1109/MPRV.2006.18

Adrian Bangerter and Daniel M. Oppenheimer. 2006. Accuracy in detecting referents of pointing gestures unaccompanied by language. *Gesture* 6, 1, 85–102.

Simon Baron-Cohen. 1995. The eye direction detector (EDD) and the shared attention mechanism (SAM): two cases for evolutionary psychology. In Chris Moore, Philip J. Dunham, and Phil Dunham, editors, *Joint Attention: Its Origin and Role in Development*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Thomas Baudel and Michel Beaudouin-Lafon. 1993. Charade: Remote control of objects using free-hand gestures. *Communications of the ACM* 36, 7, 28–35. http://doi.org/10.1145/159544.159562

Michel Beaudouin-Lafon. 2000. Instrumental interaction: an interaction model for designing post-WIMP user interfaces. In *Proceedings of the 18th conference on Human Factors in Computing Systems – CHI '00*. ACM Press, New York, NY, USA, 446–453. http://doi.org/10.1145/332040.332473

Michel Beaudouin-Lafon, Katrine Ravn, Anne Ratzer, Søren Christensen, Kurt Jensen, Wendy E. Mackay, Peter Andersen, Paul Janecek, Mads Jensen, Michael Lassen, Kasper Lund, Kjeld Mortensen, and Stephanie Munck. 2001. CPN/tools: Revisiting the desktop metaphor with post-WIMP interaction techniques. In *Extended abstracts of the 19th conference on Human Factors in Computing Systems –- CHI EA'01*. ACM Press, New York, NY, USA, 11–12. http://doi.org/10.1145/634067.634076

Michel Beaudouin-Lafon. 2004. Designing interaction, not interfaces. In *Proceedings of the 7th conference on Advanced Visual Interfaces – AVI '04*. ACM Press, New York, NY, USA, 15–22. http://doi.org/10.1145/989863.989865

Steve Benford and Lennart Fahlén. 1993. A spatial model of interaction in large virtual environments. In *Proceedings of the 3rd European Conference on Computer-Supported*

*Cooperative Work – ECSCW '93*. Berlin / Heidelberg, Germany, 109–124. http://doi.org/10.1007/978-94-011-2094-4_8

Walter Benenson, John W. Harris, Horst Stöcker, and Holger Lutz, editors. 2002. Handbook of Physics. Springer-Verlag New York, New York, NY, USA.

Niels Ole Bernsen. 1997. Defining a taxonomy of output modalities from an HCI perspective. *Computer Standards & Interfaces* 18, 6-7, 537–553. http://doi.org/10.1016/S0920-5489(97)00018-4

Sian L. Beilock, Sarah A. Wierenga, and Thomas H. Carr. 2010. Expertise, attention, and memory in sensorimotor skill execution: Impact of novel task constraints on dual-task performance and episodic memory. *The Quarterly journal of experimental psychology, Section A: Human experimental psychology* 55, 4, 1211–1240. http://doi.org/10.1080/02724980244000170

Roland M. Biedert. 2000. Contribution of the three levels of nervous system motor control: spinal cord, lower brain, cerebral cortex. In Scott M. Lephart and Freddie H. Fu, editors, *Proprioception and neuromuscular control in joint stability*. Human Kinetics, Champaign, IL, USA.

Benjamin Biguer, Claude Prablanc, and Marc Jeannerod. 1984. The contribution of coordinated eye and head movements in hand pointing accuracy. *Experimental Brain Research* 55, 3, 462–469. http://doi.org/10.1007/BF00235277

Renaud Blanch, Yves Guiard, and Michel Beaudouin-Lafon. 2004. Semantic pointing. In *Proceedings of the 22nd conference on Human Factors in Computing Systems – CHI '04*. ACM Press, New York, NY, USA, 519–526. http://doi.org/10.1145/985692.985758

Richard A. Bolt. 1980. Put-that-there: Voice and gesture at the graphics interface. In *Proceedings of the 7th conference on Computer Graphics and Interactive Techniques – GRAPH '80*. ACM Press, New York, NY, USA, 262–270. http://doi.org/10.1145/965105.807503

David L. Brock. 2001. The Electronic Product Code (EPC). Cambridge, MA, USA, 2001.

Doug A. Bowman, Chadwick A. Wingrave, Joshua M. Campbell, Vinh Q. Ly, and Christopher J. Rhoton. 2002. Novel uses of Pinch Gloves for virtual environment interaction techniques. *Virtual Reality* 6, 3, 122–129. http://doi.org/10.1007/s100550200013

Barry Brumitt and JJ Cadiz. 2001. Let there be light: Examining interfaces for homes of the future. In *Proceedings of the 8th conference on Human-Computer Interaction – INTERACT '01*. IOS Press, Amsterdam, The Netherlands, 375–382.

Robert P. Burton and Ivan E. Sutherland. 1974. Twinkle box. In *Proceedings of the conference of the American Federation of Information Processing Societies – AFIPS '74*. ACM Press, New York, NY, USA, 513–520. http://doi.org/10.1145/1500175.1500278

Xiang Cao and Ravin Balakrishnan. 2003. VisionWand: Interaction techniques for large displays using a passive wand tracked in 3D. In *Proceedings of the 16th symposium on User Interface Software and Technology – UIST' 03*. ACM Press, New York, NY, USA, 173–182. http://doi.org/10.1145/964696.964716

Maurizio Caon, Yong Yue, Julien Tscherrig, Elena Mugellini, and Omar Abou Khaled. 2011. Context-aware 3D gesture interaction based on multiple Kinects. In *Proceedings of the 1st conference on Ambient Computing, Applications, Services and Technologies – AMBIENT '11*. IARIA. 7–12.

Stuart K. Card, Thomas P. Moran, and Allen Newell. 1983. The psychology of human-computer interaction. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Noam Chomsky. 1965. Aspects of the theory of syntax. The MIT Press, Cambridge, MA, USA.

Marcus Tullius Cicero. 1988. De oratore. Harvard University Press, Cambridge, MA, USA.

Anthony A. Clarke. 1986. A three-level human-computer interface model. *International Journal of Man-Machine Studies* 24, 6, 503–517. http://doi.org/10.1016/S0020-7373(86)80006-2

Andy Cockburn and Bruce McKenzie. 2001. 3D or not 3D?: Evaluating the effect of the third dimension in a document management system. In *Proceedings of the 19th conference on Human Factors in Computing Systems – CHI '01*. ACM Press, New York, NY, USA, 434–441. http://doi.org/10.1145/365024.365309

Andy Cockburn and Bruce McKenzie. 2002. Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In *Proceedings of the 20th conference on Human Factors in Computing Systems – CHI '02*. ACM Press, New York, NY, USA, 203–210. http://doi.org/10.1145/503376.503413

Andy Cockburn, Philip Quinn, Carl Gutwin, Gonzalo Ramos, and Julian Looser. 2011. Air pointing: Design and evaluation of spatial target acquisition with and without visual feedback. *International Journal of Human-Computer Studies* 69, 6, 401–414. http://doi.org/10.1016/j.ijhcs.2011.02.005

Neal J. Cohen and Larry R. Squire. 1980. Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science* 210, 4466, 207–210. http://doi.org/10.1126/science.7414331

Gabe Cohn, Daniel Morris, Shwetak N. Patel, and Desney S. Tan. 2012. Humantenna: using the body as an antenna for real-time whole-body interaction. In *Proceedings of the 30th conference on Human Factors in Computing Systems – CHI '12*. ACM Press, New York, NY, USA, 1901–1910. http://doi.org/10.1145/2207676.2208330

Patricia Conti and Daniel Beaubaton. 1980. Role of structured visual field and visual reafference in accuracy of pointing movements. *Perceptual and Motor Skills* 50, 1, 239–244. http://doi.org/10.2466/pms.1980.50.1.239

Diane J. Cook and Sajal K. Das. 2004. Smart environments: technology, protocols and applications. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Marc K. D. Coomans and Harry J. P. Timmermans. 1997. Towards a taxonomy of virtual reality user interfaces. In *Proceedings of the 1997 IEEE Conference on Information Visualization*. IEEE, New York, NY, USA, 279–284. http://doi.org/10.1109/IV.1997.626531

Andy Crabtree, Tom Rodden, Terry Hemmings, and Steve Benford. 2003. Finding a Place for UbiComp in the Home. In *Proceedings of the 5th conference on Ubiquitous Computing - UbiComp '03*. Springer-Verlag, Berlin / Heidelberg, Germany, 208–226. http://doi.org/10.1007/978-3-540-39653-6_17

Andy Crabtree and Tom Rodden. 2004. Domestic routines and design for the home. *Computer Supported Cooperative Work* 13, 2, 191–220. http://doi.org/10.1023/B:COSU.0000045712.26840.a4

Alan Dix. 1994. Computer Supported Cooperative Work: A framework. In Duska Rosenberg and Christopher Hutchison, editors, *Design Issues in CSCW*. Springer-Verlag London, London, UK. http://doi.org/26.10.1007/978-1-4471-2029-2_2

Paul Dourish and Victoria Bellotti. 1992. Awareness and coordination in shared workspaces. In *Proceedings of the 4th conference on Computer-supported cooperative work – CSCW '92*. ACM Press, New York, NY, USA, 107–114. http://doi.org/10.1145/143457.143468

Ken Ducatel, Marc Bogdanowicz, Fabiana Scapolo, Jeroen A. J. Leijten, and Jean-Claude Burgelman. 2001. Scenarios for ambient intelligence in 2010. Institute of Prospective Technological Studies, Seville, Spain.

Yadin Dudai, Henry L. Roediger, and Endel Tulving. 2007. Memory concepts. In Henry L. Roediger, Yadin Dudai, and Susan M. Fitzpatrick, editors, *Science of Memory: Concepts*. Oxford University Press, Oxford, UK.

Susan T. Dumais and William P. Jones. 1985. A comparison of symbolic and spatial filing. In *Proceedings of the 3rd conference on Human Factors in Computing Systems – CHI '85*. ACM Press, New York, NY, USA, 127–130. http://doi.org/10.1145/1165385.317479

Tilak Dutta. 2012. Evaluation of the Kinect™ sensor for 3-D kinematic measurement in the workplace. *Applied Ergonomics* 43, 4, 645–9. http://doi.org/10.1016/j.apergo.2011.09.011

W. Keith Edwards and Rebecca E. Grinter. 2001. At home with ubiquitous computing: seven challenges. In *Proceedings of the 3rd conference on Ubiquitous Computing – UbiComp '01*. Springer-Verlag, Berlin / Heidelberg, Germany, 256–272. http://doi.org/10.1007/3-540-45427-6_22

David Efron. 1941. Gesture and environment. Ph.D. Dissertation. Columbia University, New York, NY, USA

Paul Ekman and Wallace V. Friesen. 1981. The repertoire of nonverbal behavior. In Adam Kendon, Thomas A. Sebeok, and Jean Umiker-Sebeok, editors, *Nonverbal communication, interaction, and gesture*. Mouton Publishers, The Hague, The Netherlands, 1981, 57–105.

Lorin J. Elias and Deborah M. Saucier. 2006. Neuropsychology: Clinical and Experimental Foundations. Pearson Education, Boston, MA, US.

Mica R. Endsley. 1988. Design and Evaluation for Situation Awareness Enhancement. *Human Factors: Annual Meeting of the Human Factors and Ergonomics Society* 32, 2, 97–101. http://doi.org/10.1177/154193128803200221

Mica R. Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1, 32–64. http://doi.org/10.1518/001872095779049543

Douglas C. Engelbart. 1970. X-Y position indicator for a display system. No. US3541541 A, United States of America.

Thomas Erber and George M. Hockney. 2007. Complex systems: equilibrium configurations of N equal charges on a sphere ($2 \leq N \geq 112$). *Advances in Chemical Physics* 98, 495–594. http://doi.org/10.1002/9780470141571.ch5

Paul M. Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47, 6, 381–391.

Paul M. Fitts and Michael I. Posner. 1967. Human Performance. Greenwood Press, Westpoint, CT, USA.

George W. Fitzmaurice. 1993. Situated information spaces and spatially aware palmtop computers. *Communications of the ACM* 36, 7, 39–49. http://doi.org/10.1145/159544.159566

George W. Fitzmaurice, Hiroshi Ishii, and William A. S. Buxton. 1995. Bricks: Laying the foundations for graspable user interfaces. In *Proceedings of the 13th conference on Human Factors in Computing Systems – CHI '95*. ACM Press, New York, NY, USA, 442–449. http://doi.org/10.1145/223904.223964

Douglas Galasko, David Bennett, Mary Sano, Chris Ernesto, Ronald Thomas, Michael Grundman, and Steven Ferris. 1997. An inventory to assess activities of daily living for clinical trials in Alzheimer's disease. *Alzheimer Disease & Associated Disorders* 11, 2, S33–S39.

Neil Gershenfeld, Raffi Krikorian, and Danny Cohen. 2004. The Internet of Things. *Scientific American* 291, 4, 76–81. http://doi.org/10.1038/scientificamerican1004-76

Tony Gillie and Donald Broadbent. 1989. *What makes interruptions disruptive? A study of length, similarity, and complexity*. Psychological Research 50, 4, 243–250.

James Gordon, Maria F. Ghilardi, and Claude Ghez. 1995. Impairments of reaching movements in patients without proprioception. I. Spatial errors. *Journal of Neurophysiology* 73, 1, 347–360.

Sean Gustafson, Daniel Bierwirth, and Patrick Baudisch. 2010. Imaginary interfaces: Spatial interaction with empty hands and without visual feedback. In *Proceedings of the 23rd symposium on User Interface Software and Technology – UIST '10*. ACM Press, New York, NY, USA, 3–12. http://doi.org/10.1145/1866029.1866033

Tiago Guerreiro, Ricardo Gamboa, and Joaquim A. Jorge. 2007. Mnemonical body shortcuts for interacting with mobile devices. In *Proceedings of the 7th international Gesture Workshop – GW '07*. 261–271. http://doi.org/10.1007/978-3-540-92865-2_29

Carl Gutwin and Saul Greenberg. 1996. Workspace awareness for groupware. In *Proceedings of the 14th conference on Human Factors in Computing Systems – CHI '96*. ACM Press, New York, NY, USA, 208–209. http://doi.org/10.1145/257089.257284

Carl Gutwin and Saul Greenberg. 1998. Design for individuals, design for groups: tradeoffs between power and workspace awareness. In *Proceedings of the 7th conference on Computer Supported Cooperative Work – CSCW '98*. ACM Press, New York, NY, USA, 207–216. http://doi.org/10.1145/289444.289495

Carl Gutwin and Saul Greenberg. 2002. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work* 11, 3-4, 411–446. http://doi.org/10.1023/A:1021271517844

Carl Gutwin, Andy Cockburn, Joey Scarr, Sylvain Malacria, and Scott C. Olson. 2014. Faster command selection on tablets with FastTap. In *Proceedings of the 32nd conference on Human Factors in Computing Systems – CHI '14*. ACM Press, New York, NY, USA, 2617–2626. http://doi.org/10.1145/2556288.2557136

Elizabeth Hanna and Andrew N. Meltzoff. 1993. Peer imitation by toddlers in laboratory, home, and day-care contexts: Implications for social learning and memory. *Developmental Psychology* 29, 4, 701–710. http://doi.org/10.1037/0012-1649.29.4.701

Walter J. Hendelman. 2005. Atlas of functional neuroanatomy. CRC Press, Boca Raton, FL, USA.

Ken Hinckley, Randy Pausch, John C. Goble, and Neal F. Kassell. 1994. Passive real-world interface props for neurosurgical visualization. In *Proceedings of the 12th conference on Human Factors in Computing Systems – CHI '94*. ACM Press, New York, NY, USA, 452–458. http://doi.org/10.1145/191666.191821

John J. Hopfield. 1982. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences* 79, 8, 2554–2558. http://doi.org/10.1073/pnas.79.8.2554

Juan Pablo Hourcade and Natasha E. Bullock-Rest. 2012. How small can you go?: Analyzing the effect of visual angle in pointing tasks. In *Proceedings of the 30th conference on Human Factors in Computing Systems – CHI '12*. ACM Press, New York, NY, USA, 213–216. http://doi.org/10.1145/2207676.2207706

Hiroshi Ishii and Brygg Ullmer. 1997. Tangible bits: Towards seamless interfaces between people, bits and atoms. In *Proceedings of the 15th conference on Human Factors in Computing Systems – CHI '97*. ACM Press, New York, NY, USA, 234–241. http://doi.org/10.1145/258549.258715

Neal F. Johnson. 1969. Chunking: Associative chaining versus coding. *Journal of Verbal Learning and Verbal Behavior* 8, 6, 725–731. http://doi.org/10.1016/S0022-5371(69)80036-8

William P. Jones. 2007. Personal Information Management. *Annual Review of Information Science and Technology* 41, 1, 453–504. http://doi.org/10.1002/aris.2007.1440410117

Maria Karam and Monica C. Schraefel. 2005. A taxonomy of gestures in human computer interactions. Southampton, UK.

Charles C. Kemp, Cressel D. Anderson, Hai Nguyen, Alexander J. Trevor, and Zhe Xu. 2008. A point-and-click interface for the real world: Laser designation of objects for mobile manipulation. In *Proceedings of the 3rd conference on Human-Robot Interaction – HRI '08*. ACM Press, New York, NY, USA, 241–248. http://doi.org/10.1145/1349822.1349854

Teuvo Kohonen. 1984. Self-Organization and Associative Memory. Springer-Verlag, Berlin / Heidelberg, Germany.

Régis A. Kopper, Doug A. Bowman, Mara G. Silva, and Ryan P. McMahan. 2010. A human motor behavior model for distal pointing tasks. *International Journal of Human-Computer Studies* 68, 10, 603–615. http://doi.org/10.1016/j.ijhcs.2010.05.001

Arthur F. Kramer and Andrew Jacobson. 1991. Perceptual organization and focused attention: The role of objects and proximity in visual processing. *Perception & Psychophysics* 50, 3, 267–284. http://doi.org/10.3758/BF03206750

Myron W. Krueger, Thomas Gionfriddo, and Katrin Hinrichsen. 1985. VIDEOPLACE: An artificial reality. In *Proceedings of the 4th conference on Human Factors in Computing Systems – CHI '85*. ACM Press, New York, NY, USA, 35–40. http://doi.org/10.1145/1165385.317463

Christine Kühnel, Tilo Westermann, Fabian Hemmert, Sven Kratz, Alexander Müller, and Sebastian Möller. 2011. I'm home: Defining and evaluating a gesture set for smart-home control. *International Journal of Human-Computer Studies* 69, 11, 693–704. http://doi.org/10.1016/j.ijhcs.2011.04.005

Gordon Kurtenbach, Abigail Sellen, and William Buxton. 1993. An Empirical Evaluation of Some Articulatory and Cognitive Aspects of Marking Menus. *Human-Computer Interaction* 8, 1, 1–23. http://doi.org/10.1207/s15327051hci0801_1

Gordon Kurtenbach and William Buxton. 1994. User learning and performance with marking menus. In *Proceeding of the 12th conference on Human Factors in Computing Systems – CHI '94*. ACM Press, New York, NY, USA, 258–264. http://doi.org/10.1145/191666.191759

James A. Landay and Gaetano Borriello. 2003. Design patterns for ubiquitous computing. *Computer* 36, 8, 93–95. . http://doi.org/10.1109/MC.2003.1220589

Mark W. Lansdale. 1988. The psychology of personal information management. *Applied Ergonomics* 19, 1, 55–66. http://doi.org/10.1016/0003-6870(88)90199-8

Joseph J. LaViola. 2014. An introduction to 3D gestural interfaces. In *Courses of the 41st conference on Computer Graphics and Interactive Techniques – GRAPH '14*. ACM Press, New York, NY, USA, 1–42. http://doi.org/10.1145/2614028.2615424

Maria Lehnung, Bernd Leplow, Vegard Oksendal Haaland, Maximilian Mehdorn, and Roman Ferstl. 2003. Pointing accuracy in children is dependent on age, sex and experience. *Journal of Environmental Psychology* 23, 4, 419–425. http://doi.org/10.1016/S0272-4944(02)00084-1

Jakob Leitner and Michael Haller. 2011. Geckos: combining magnets and pressure images to enable new tangible-object design and interaction. In *Proceedings of the 29th conference on Human factors in computing systems – CHI '11*. ACM Press, New York, NY, USA, 2985–2994. http://doi.org/10.1145/1978942.1979385

Scott M. Lephart, Bryan L. Rieman, and Freddie H. Fu. 2000. Introduction to the sensorimotor system. In Scott M. Lephart and Freddie H. Fu, editors, *Proprioception and neuromuscular control in joint stability*. Human Kinetics, Champaign, IL, USA.

Frank Chun Yat Li, David Dearman, and Khai N. Truong. 2009. Virtual shelves: Interactions with orientation aware devices. In *Proceedings of the 22nd symposium on User Interface*

*Software and Technology – UIST '09*. ACM Press, New York, NY, USA, 125–128. http://doi.org/10.1145/1622176.1622200

Gordon D. Logan. 1988. Toward an instance theory of automatization. *Psychological Review* 95, 4, 492–527. http://doi.org/10.1037/0033-295X.95.4.492

Steven J. Luck and Andrew Hollingworth, editors. 2008. Visual Memory. Oxford University Press, New York, NY, USA.

Richard F. Lyon. 1981. The optical mouse, and an architectural methodology for smart digital sensors. In Hsiang-Tsung Kung, Bob Sproull, and Guy Steele, editors, *VLSI Systems and Computations*. Springer-Verlag, Berlin / Heidelberg, Germany.

Hongshen Ma and Joseph A. Paradiso. 2002. The FindIT flashlight: Responsive tagging based on optically triggered microprocessor wakeup. In *Proceedings of the 4th conference on Ubiquitous Computing – UbiComp '02*. Springer-Verlag, Berlin / Heidelberg, Germany, 655–662. http://doi.org/10.1007/3-540-45809-3_12

Ian S. MacKenzie. 1992. Fitts' Law as a Research and Design Tool in Human-Computer Interaction. *Human-Computer Interaction* 7, 1, 91–139. http://doi.org/10.1207/s15327051hci0701_3

Ian S. MacKenzie and Shaidah Jusoh. 2001. An evaluation of two input devices for remote pointing. In *Proceedings of the 8th IFIP Working Conference on Engineering for Human-Computer Interaction – EHCI '01*. 235–250. http://doi.org/10.1007/3-540-45348-2_21

Donald G. MacKay. 1982. The problems of flexibility, fluency, and speed–accuracy trade-off in skilled behavior. *Psychological Review* 89, 5, 483–506. http://doi.org/10.1037/0033-295X.89.5.483

Thomas W. Malone. 1983. How do people organize their desks?: Implications for the design of office information systems. *ACM Transactions on Information Systems* 1, 1, 99–112. http://doi.org/10.1145/357423.357430

Teresa Marrin. 1997. Possibilities for the digital baton as a general-purpose gestural interface. In *Extended abstracts of the 15th conference on Human Factors in Computing Systems – CHI EA '97*. ACM Press, New York, NY, USA, 311. http://doi.org/10.1145/1120212.1120409

Ronald G. Marteniuk, Christine L. Mackenzie, Marc Jeannerod, Sylvie Athenes, and Claude Dugas. 1987. Constraints on human arm movement trajectories. *Canadian Journal of Psychology* 43, 1, 365–378. http://doi.org/10.1037/h0084157

David McGookin and Stephen Brewster. 2011. Investigating Phicon feedback in non-visual tangible user interfaces. In *Work-in-progress of the 29th conference on Human factors in computing systems – CHI WIP '11*. ACM Press, New York, NY, USA, 1543–1548. http://doi.org/10.1145/1979742.1979805

David McNeill. 1992. Hand and mind: What gestures reveal about thought. University of Chicago Press, Chicago, IL, USA,.

David McNeill. 2005. Gesture and thought. University of Chicago Press, Chicago, IL, USA.

Paul Milgram and Fumio Kishino. 1994. A taxonomy of mixed reality visual displays. *Transaction on Information and Systems* 77, 12, 1321–1329.

Pranav Mistry, Pattie Maes, and Liyan Chang. 2009. WUW - wear Ur world: A wearable gestural interface. In *Extended abstracts of the 28th conference on Human Factors in Computing Systems – CHI EA '09*. ACM Press, New York, NY, USA, 4111. http://doi.org/10.1145/1520340.1520626

Charles W. Morris. 1971. Writings on the general theory of signs. The Hague, The Netherlands.

Daniel Morris, T. Scott Saponas, and Desney S. Tan. 2010. Emerging input technologies for always-available mobile interaction. *Foundations and Trends in Human–Computer Interaction* 4, 4, 245–316. http://doi.org/10.1561/1100000023

Brad A. Myers, Rishi Bhatnagar, Jeffrey Nichols, et al. 2002. Interacting at a distance: Measuring the performance of laser pointers and other devices. In *Proceedings of the 20th conference on Human Factors in Computing Systems – CHI '02*. ACM Press, New York, NY, USA, 33–40. http://doi.org/10.1145/503376.503383

Suzanne Nalbantian. 2011. Memory and imagination in romantic fiction. In Suzanne Nalbantian, Paul M. Matthews, and James L. McClelland, editors, *The Memory Process: Neuroscientific and Humanistic Perspectives*. The MIT Press, Cambridge, MA, USA.

Katherine Nelson. 1988. Constraints on word learning? *Cognitive Development* 3, 3, 221–246. http://doi.org/10.1016/0885-2014(88)90010-X

Allen Newell and Paul S. Rosenbloom. 1981. Mechanisms of skill acquisition and the law of practice. In John R. Anderson, editor, *Cognitive skills and their acquisition*. Lawrence Erlbaum Associates, Pittsburgh, PA, USA, 1–56.

Donald A. Norman. 1993. Things that make us smart: defending human attributes in the age of the machine. Perseus Books Group, New York, NY, USA.

Winfried Nöth. 1995. Handbook of semiotics. Indiana University Press, Bloomington, IN, USA.

Jon O'Brien and Tom Rodden. 1997. Interactive systems in domestic environments. In *Proceedings of the 2nd conference on Designing Interactive Systems – DIS '97*. ACM Press, New York, NY, USA, 247–259. http://doi.org/10.1145/263552.263617

Ji-Young Oh and Wolfgang Stuerzlinger. 2002. Laser pointers as collaborative pointing devices. In *Proceedings of the 21st conference on Graphics Interfaces – GI '02*. ACM Press, New York, NY, USA.

Kenton O'Hara, Gerardo Gonzalez, Graeme Penney, Abigail Sellen, Robert Corish, Helena Mentis, Andreas Varnavas, Antonio Criminisi, Mark Rouncefield, Neville Dastur, and Tom Carrell. 2014. Interactional order and constructed ways of seeing with touchless imaging systems in surgery. *Computer Supported Cooperative Work* 23, 3, 299–337. http://doi.org/10.1007/s10606-014-9203-4

Dan R. Olsen and Travis Nielsen. 2001. Laser pointer interaction. In *Proceedings of the 19th conference on Human Factors in Computing Systems – CHI '01*. ACM Press, New York, NY, USA, 17–22. http://doi.org/10.1145/365024.365030

Allan Paivio. 1971. Imagery and verbal processes. Holt, Rinehart, & Winston, New York, NY, USA.

Thomas J. Palmeri and Michael J. Tarr. 2008. Visual object perception and long-term memory. In Steven J. Luck and Andrew Hollingworth, editors, *Visual memory*. Oxford University Press, Oxford, UK.

Shwetak N. Patel and Gregory D. Abowd. 2003. A 2-way laser-assisted selection scheme for handhelds in a physical environment. In *Proceedings of the 3th conference on Pervasive and Ubiquitous Computing – UbiComp '03*. Springer-Verlag, Berlin / Heidelberg, Germany, 200–207. http://doi.org/10.1007/b93949

Simon T. Perrault, Eric Lecolinet, Yoann P. Bourse, Shengdong Zhao, and Yves Guiard. 2015. Physical Loci: leveraging spatial, object and semantic memory for command selection. In *Proceedings of the 33rd conference on Human Factors in Computing Systems – CHI '15*. ACM Press, New York, NY, USA, 299–308. http://doi.org/10.1145/2702123.2702126

Leo Postman. 1962. The effects of language habits on the acquisition and retention of verbal associations. *Journal of Experimental Psychology* 64, 1, 7–19. http://doi.org/10.1037/h0041123

Leo Postman, Janat Fraser, and Sheila Burns. 1968. Unit-sequence facilitation in recall. *Journal of Verbal Learning and Verbal Behavior* 7, 1, 217–224. http://doi.org/10.1016/S0022-5371(68)80192-6

Daniel J. Povinelli, James E. Reaux, Donna T. Bierschwale, Ashley D. Allain, and Bridgett B. Simon. 1997. Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees. *Cognitive Development* 12, 4, 423–461. http://doi.org/10.1016/S0885-2014(97)90017-4

Qifan Pu, Sidhant Gupta, Shyam Gollakota, and Shwetak N. Patel. 2013. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th conference on Mobile Computing and Networking – MobiCom '13*. ACM Press, New York, NY, USA, 27–38. http://doi.org/10.1145/2500423.2500436

Zenon W. Pylyshyn and Neil M. Agnew. 1963. The validity of anxiety and drive scales and their relation to global self-ratings. *The Canadian Psychologist* 4, 2, 42–50.

Jun Rekimoto and Katashi Nagao. 1995. The world through the computer. In *Proceedings of the 8th symposium on User Interface and Software Technology – UIST '95*. ACM Press, New York, NY, USA, 29–36. http://doi.org/10.1145/215585.215639

Jun Rekimoto and Masanori Saitoh. 1999. Augmented surfaces: A spatially continuous work space for hybrid computing environments. In *Proceedings of the 17th conference on Human Factors in Computing Systems – CHI '99*. ACM Press, New York, NY, USA, 378–385. http://doi.org/10.1145/302979.303113

Hongliang Ren, Wei Liu, and Andy Lim. 2013. Marker-Based Surgical Instrument Tracking Using Dual Kinect Sensors. *IEEE Transactions on Automation Science and Engineering* 11, 3, 1–4. http://doi.org/10.1109/TASE.2013.2283775

Harriet L. Rheingold and Kaye V. Cook. 1975. The contents of boys' and girls' rooms as an index of parents' behavior. *Child development* 46, 2, 459–463.

Julie Rico and Stephen Brewster. 2010. Usable gestures for mobile interfaces: evaluating social acceptability. In *Proceedings of the 28th conference on Human Factors in Computing Systems – CHI '10*. New York, NY, USA, 887–896. http://doi.org/10.1145/1753326.1753458

Lawrence G. Roberts. 1966. The Lincoln WAND. In *Proceedings of the conference of the American Federation of Information Processing Societies – AFIPS '66*. ACM Press, New York, NY, USA, 223–227. http://doi.org/10.1145/1464291.1464317

George G. Robertson, Mary Czerwinski, Kevin Larson, Daniel C. Robbins, David Thiel, and Maarten van Dantzich. 1998. Data mountain: Using spatial memory for document management. In *Proceedings of the 11th symposium on User Interface Software and Technology – UIST '98*. ACM Press, New York, NY, USA, 153–162. http://doi.org/10.1145/288392.288596

Henry L. Roediger. 1990. Implicit memory: A commentary. *Bulletin of the Psychonomic Society* 28, 4, 373–380. http://doi.org/10.3758/BF03334044

Tony Salvador, Jean Scholtz, and James Larson. 1996. The Denver model for groupware design. *ACM SIGCHI Bulletin* 28, 1, 52–58. http://doi.org/10.1145/249170.249185

Scott Saponas, Desney S. Tan, Dan Morris, and Ravin Balakrishnan. 2008. Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces. In *Proceeding of the 26th conference on Human factors in computing systems – CHI '08*. ACM Press, New York, NY, USA, 515–524. http://doi.org/10.1145/1357054.1357138

Mahadev Satyanarayanan. 2001. Pervasive computing: vision and challenges. *IEEE Personal Communications* 8, 4, 10–17. http://doi.org/10.1109/98.943998

Joey Scarr, Andy Cockburn, Carl Gutwin, and Andrea Bunt. 2012. Improving command selection with CommandMaps. In *Proceedings of the 30th conference on Human Factors in Computing Systems – CHI '12*. ACM Press, New York, NY, USA, 257–266. http://doi.org/10.1145/2207676.2207713

Daniel L. Schacter and Endel Tulving. 1994. What are the memory systems of 1994? In Daniel L. Schacter and Endel Tulving, editors, *Memory Systems 1994*. The MIT Press, Cambridge, MA, USA.

Chris L. Schmidt. 1995. Adult understanding of spontaneous attention-directing events: what does gesture contribute? In Chris Moore, Philip J. Dunham, and Phil Dunham, editors, *Joint Attention: Its Origin and Role in Development*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Richard A. Schmidt. 1975. A schema theory of discrete motor skill learning. *Psychological Review* 84, 2, 225–260. http://doi.org/10.1037/h0076770

Richard A. Schmidt, Howard Zelaznik, Brian Hawkins, James S. Frank, and John T. Quinn. 1979. Motor-output variability: a theory for the accuracy of rapid motor acts. *Psychological Review* 86, 5, 415–451.

Richard A. Schmidt and Timothy D. Lee. 2005. Motor control and learning: A behavioral emphasis. Human Kinetics, Champaign, IL, USA.

L. D. Segal. 1994. Actions speak louder than words: how pilots use nonverbal information for crew communications. In *Proceedings of the 8th Human Factors Society Annual Meeting – HFES '94*. SAGE Publications, 21–40. http://doi.org/10.1177/154193129403800106

Jakub Segen and Senthil Kumar. 1998. Human-computer interaction using gesture recognition and 3D hand tracking. In *Proceedings of the 15th International Conference on Image Processing – ICIP '98*. IEEE, 188–192. http://doi.org/10.1109/ICIP.1998.727164

Charles S. Sherrington. 1906. The integrative action of the nervous system. Yale University Press, New Haven, CT, USA.

Ben Shneiderman. 1997. Designing the user interface: strategies for effective human-computer interaction. Addison Wesley Longman, Reading, MA, USA.

Anna Shumway-Cook and Marjorie H. Woollacott. 2001. Motor control: theory and practical applications. Lippincott Williams & Wilkins, Baltimore, MA, USA.

Michael W. Smith, Joseph Sharit, and Sara J. Czaja. 1999. Aging, motor control, and the performance of computer mouse tasks. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 41, 3, 389–396. http://doi.org/10.1518/001872099779611102

John F. Soechting and Francesco Lacquaniti. 1981. Invariant characteristics of a pointing movement in man. *The Journal of Neuroscience* 1, 7, 710–720.

Larry R. Squire and Stuart M. Zola. 1996. Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences* 93, 24, 13515–13522. http://doi.org/10.1073/pnas.93.24.13515

Jürgen Streeck. 1993. Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs* 60, 4, 275–299. http://doi.org/10.1080/03637759309376314

Margaret-Anne D. Storey, Davor Čubranić, and Daniel M. German. 2005. On the use of visualization to support awareness of human activities in software development. In *Proceedings of the 2nd symposium on Software Visualization – SoftVis '05*. ACM Press, New York, NY, USA, 193–202. http://doi.org/10.1145/1056018.1056045

Steven Strachan, Roderick Murray-Smith, and Sile O'Modhrain. 2007. BodySpace: Inferring body pose for natural control of a music player. In *Extended abstracts of the 25th conference on Human Factors in Computing Systems – CHI EA'07*. ACM Press, New York, NY, USA, 2001–2006. http://doi.org/10.1145/1240866.1240939

Norbert Streitz, Jörg Haake, Jörg Hannemann, Andreas Lemke, Wolfgang Schuler, Helge Schütt, and Manfred Thüring. 1992. SEPIA: a cooperative hypermedia authoring environment. In *Proceedings of the 5th European Conference on Hypertext – ECHT '92*. ACM Press, New York, NY, USA, 11–22. http://doi.org/10.1145/168466.168479

Aimée M. Surprenant and Ian Neath. 2009. Principles of memory. Psychology Press, New York, NY, USA.

Ivan E. Sutherland. 1968. A head-mounted three dimensional display. In *Proceedings of the conference of the American Federation of Information Processing Societies – AFIPS '68*. ACM Press, New York, NY, USA, 757–764. http://doi.org/10.1145/1476589.1476686

Edward C. Tolman. 1948. Cognitive maps in rats and men. *Psychological Review* 55, 4, 189–208. http://doi.org/10.1037/h0061626

Peter Tolmie, James Pycock, Tim Diggins, Allan MacLean, and Alain Karsenty. 2002. Unremarkable computing. In *Proceeding of the 20th conference on Human Factors in Computing Systems – CHI '02*. ACM Press, New York, NY, USA, 399–406. http://doi.org/10.1145/503376.503448

Michael Tomasello. 1995. Joint attention as social cognition. In Chris Moore, Philip J. Dunham, and Phil Dunham, editors, *Joint Attention: Its Origin and Role in Development*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

László F. Tóth. 1943. Über eine Abschätzung des kürzesten Abstandes zweier Punkte eines auf einer Kugelfläche liegenden Punktsystems. *Jahresbericht der Deutschen Mathematiker-Vereinigung* 53, 66–68.

Angela K. Troyer, Morris Moscovitch, and Gordon Winocur. 1997. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology* 11, 1, 138–146. http://doi.org/10.1037//0894-4105.11.1.138

Endel Tulving. 1985. How many memory systems are there? *American Psychologist* 40, 4, 385–398. http://doi.org/10.1037/0003-066X.40.4.385

Paul D. Tynan and Robert Sekuler.1982. Motion processing in peripheral vision: Reaction time and perceived velocity. *Vision Research* 22, 1, 61–68. http://doi.org/10.1016/0042-6989(82)90167-5

John Vardalas. 1994. From DATAR to the FP-6000: Technological change in a Canadian industrial context. *IEEE Annals of the History of Computing* 16, 2, 20–30. http://doi.org/10.1109/85.279228

Daniel Vogel and Ravin Balakrishnan. 2005. Distant freehand pointing and clicking on very large, high resolution displays. In *Proceedings of the 18th symposium on User Interface Software and Technology – UIST '05*. ACM Press, New York, NY, USA, 33–42. http://doi.org/10.1145/1095034.1095041

Nicholas J. Wade and Michael Swanston. 1991. Visual Perception. Routledge, London, UK.

Denis Weaire and Tomaso Aste. 2008. The pursuit of perfect packing. Taylor & Francis Group, Boca Raton, FL, USA.

Mark Weiser. 1991. The computer for the 21st century. *Scientific American* 265, 3, 94–104. http://doi.org/10.1038/scientificamerican0991-94

Mark Weiser. 1993. Some computer science issues in ubiquitous computing. *Communications of the ACM* 36, 7, 75–84. http://doi.org/10.1145/159544.159617

Pierre Wellner. 1993. Interacting with paper on the DigitalDesk. *Communications of the ACM* 36, 7, 87–96. http://doi.org/10.1145/159544.159630

Pierre Wellner, Wendy Mackay, and Rich Gold. 1993. Back to the real world. *Communications of the ACM* 36, 7, 24–27. http://doi.org/10.1145/159544.159555

David J. Willshaw, Oscar P. Buneman, and Hugh C. Longuet-Higgens. 1969. Non-Holographic Associative Memory. *Nature* 222, 5197, 960–962. http://doi.org/10.1038/222960a0

Andrew Wilson and Nuria Oliver. 2003. GWindows: robust stereo vision for gesture-based control of windows. In *Proceedings of the 5th International Conference on Multimodal*

*Interfaces – ICMI '03*. New York, NY, USA, 211–218. http://doi.org/10.1145/958432.958473

Andrew Wilson and Hubert Pham. 2003. Pointing in intelligent environments with the World Cursor. In *Proceedings of the 9th conference on Human-Computer Interaction – INTERACT '03*. Springer-Verlag, Berlin / Heidelberg, Germany, 495–502.

Andrew Wilson and Steven Shafer. 2003. XWand: UI for intelligent spaces. In *Proceedings of the 21st conference on Human Factors in Computing Systems – CHI '03*. ACM Press, New York, NY, USA, 545–552. http://doi.org/10.1145/642611.642706

Jong-bum Woo and Youn-kyung Lim. 2012. Clipoid: An augmentable short-distance wireless toolkit for 'accidentally smart home' environments. In *Proceedings of the 30th conference on Human Factors in Computing Systems – CHI '12*. ACM Press, New York, NY, USA, 1751–1754. http://doi.org/10.1145/2207676.2208305

Frances A. Yates. 1966. The Art of Memory. Routledge & Kegan Paul, London, UK, 1966.

Computer History Museum: The Mouse. *Computer History Museum*, 2015a. http://www.computerhistory.org/revolution/input-output/14/350

Computer History Museum: The Xerox Alto. *Computer History Museum*, 2015b. http://www.computerhistory.org/revolution/input-output/14/347

NaturalPoint OptiTrack Prime. *NaturalPoint Inc.*, 2015. http://www.naturalpoint.com/optitrack/hardware/

Vicon Bonita. *Vicon Motion Systems Ltd.*, 2015. http://www.vicon.com/products/camera-systems/bonita

# Appendix A: Glossary and Abbreviations

**C** **Control device**: a piece of hardware that records user input for interaction with a digital artifact

**D** **Device-free interaction**: interacting with a digital system without holding or touching an interaction device

**Digital artifact**: anything that can be selected through a digital system, e.g., commands, files, and bookmarks

**Digital device**: device that can be controlled remotely through a control device, e.g., TV set; digital devices are a subset of all digital artifacts

**E** **Environment**: physical space confined by walls, a floor, and a ceiling; area and height are limited to what can typically be found in domestic and office settings (see 3.3.1)

**Eyes-free interaction**: interacting with a digital system without paying visual attention to the interaction

**G** **Goal**: an overall state that people want to achieve by manipulating the environment.

**GOMS, Goal–Operator–Methods–Selection-model**: The GOMS-model is used in high-level task analysis for splitting tasks into four components: goals, operators, methods, and selection; GOMS is closely related to MHP (see 2.4.5).

**Group**: a number of people in the same environment; groups typically consist of between 2 and 20 individuals.

**H** **HCI**: Human-Computer Interaction

**HEI**, **Human-Environment Interaction**: interaction between a user and a digital system that is hidden from the user by being built into the environment; a subarea of HCI

**I** **In-place interaction**: a set of interaction techniques that use stationary interaction devices and thus require users to walk up to the device

**M** **MHP**, **Model-Human Processor**. The Model-Human Processor is used in low-level task analysis for splitting GOMS-operators in perceptual, cognitive, and motor components (see 2.4.5).

**Mid-air full-arm pointing gesture**: a gesture toward a real-world object that involves moving the entire arm (shoulder to finger) and where the arm is not supported by another body part or an object

**N** **Navigation-based interaction**: a set of interaction techniques that use flat or hierarchical on-screen menus with buttons for user interaction; selection mechanisms can be, for example, touch-based or mid-air full-arm pointing gestures

**P** **Pointing target**: the representation of a real-world selection proxy in the model of the environment. In my implementation of *Room Pointing*, the pointing target is a single vector that is located roughly at the center of the real-world object.

**Primary task**: an activity that people have to complete in order to reach a goal and that directly contributes to reaching a goal

**R** **RCAP**: Ray-Casting Air-Pointing (see 2.2.3)

**Real-world object**: a physical object (e.g., a desk) or conceptual region (e.g., a wall) in an environment

**Room Pointing**: a selection technique for HEI that is an example for room-based interaction

**Room-based interaction**: a group of selection techniques that uses mid-air full-arm pointing gestures toward real-world objects as selection proxies for selecting digital artifacts

**S** **Selection accuracy**: the percentage of correctly selected selection proxies; common measurement in HCI for evaluating interaction techniques.

**Selection mechanism**: the action or method of interacting with a digital system (e.g., pointing, direct touch)

**Selection proxy**: the digital representation of a digital artifact to the user (e.g., icon, terminal command)

**Selection speed**: the times it takes to complete a selection; common measurement in HCI for evaluating interaction techniques.

**Shared space**: an environment that is concurrently used by a group of people

**Smart environment**: environment with digital artifacts and devices that can be controlled remotely but might not have a dedicated user interface.

**Supporting task**: an activity that people have to complete in order to progress with their primary task

**System-feedback-free interaction**: interacting with a digital system without paying attention to the feedback from the system during or after the interaction

**T** **Target zone**: the area around a real-world selection proxy in which mid-air full-arm pointing gestures will select the digital artifact that is associated with the real-world object.

**Touch-based interaction**: a set of interaction techniques that use direct touch as selection mechanism

**U** **UbiComp**: Ubiquitous Computing (see 2.1.1 and 2.2.2)

**W** **WIMP**: Windows, Icons, Menus, Pointer; most common interface paradigm for graphical user interfaces on personal computers.

# Appendix B: Study Materials

## 10.1 Study 1 (Chapter 5)

**DEPARTMENT OF COMPUTER SCIENCE**
**UNIVERSITY OF SASKATCHEWAN**
**INFORMED CONSENT FORM**

UNIVERSITY OF SASKATCHEWAN

Research Project: **Gesture Observation and Recognition Study**

Investigators: Dr. Carl Gutwin, Department of Computer Science (966-8646)
Adrian Reetz, Department of Computer Science (966-2327)

This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is abo0075t and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information.

This study is concerned with detecting **performance of different selection techniques**.

The goal of the research is to **find more efficient ways for people to interact with ambient-intelligent environments**.

The session will require **60** minutes, during which you will be asked to **perform selections using four different selection techniques** in the Human-Computer Interaction Lab at the University of Saskatchewan.

At the end of the session, you will be given more information about the purpose and goals of the study, and there will be time for you to ask questions about the research. As a way of thanking you for your participation and to help compensate you for your time and any travel costs you may have incurred, you will receive a **$10** honorarium at the end of the session.

The data collected from this study will be used in articles for publication in journals and conference proceedings.

As one way of thanking you for your time, we will be pleased to make available to you a summary of the results of this study once they have been compiled (usually within two months). This summary will outline the research and discuss our findings and recommendations. This summary will be available on the HCI lab's website: **http://www.hci.usask.ca/**

All personal and identifying data will be kept confidential. Confidentiality will be preserved by using pseudonyms in any presentation of textual data in journals or at conferences. The informed consent form and all research data will be kept in a secure location under confidentiality in accordance with University policy for 5 years post publication. Do you have any questions about this aspect of the study?

**You are free to withdraw from the study at any time without penalty and without losing any advertised benefits.** Withdrawal from the study will not affect your academic status or your access to services at the university. If you withdraw, your data will be deleted from the study and destroyed. Your right to withdraw data from the study will apply until results have been disseminated, data has been pooled, etc. After this, it is possible that some form of research dissemination will have already occurred and it may not be possible to withdraw your data.

Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact:

- Dr. Carl Gutwin, Professor, Dept. of Computer Science, (306) 966-8646, gutwin@cs.usask.ca

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a participant. In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. If you have further questions about this study or your rights as a participant, please contact:

- Dr. Carl Gutwin, Professor, Dept. of Computer Science, (306) 966-8646, gutwin@cs.usask.ca
- Research Ethics Office, University of Saskatchewan, (306) 966-2975 or toll free at 888-966-2975.

Participant's signature:_____

Date:_____

Investigator's signature:_____

Date:_____

A copy of this consent form has been given to you to keep for your records and reference. This research has the ethical approval of the Research Ethics Office at the University of Saskatchewan.

# Demographics Questionnaire

**Gender:** ☐ female ☐ male ☐ other

**Age:** _____ years

**Handedness:** ☐ left ☐ right ☐ ambidextrous

**Do you own a smart phone?**

☐ no ☐ yes

**How many hours *per day* do you use your smart phone?**

☐ 0-1 hours ☐ 1-2 hours ☐ 2-3 hours ☐ 3-4 hours ☐ more than 4 hours

**How often do you carry your smart phone with you *when you are at home*?**

☐ never ☐ rarely ☐ half of the times ☐ frequently ☐ constantly

**Have you ever played a gesture-controlled video game (e.g., on Nintendo Wii or Microsoft Xbox)?**

☐ no ☐ yes

**How many hours *per week* do you play motion-controlled video games?**

☐ 0-1 hours ☐ 1-3 hours ☐ 3-6 hours ☐ 6-12 hours ☐ more than 12 hours

Pid: _____

# TLX Questionnaire

*First interaction technique: _____*

Mental Demand — How mentally demanding was the task?


Physical Demand — How physically demanding was the task?


Temporal Demand — How hurried or rushed was the pace of the task?


Performance — How successful were you in accomplishing what you were asked to do?


Effort — How hard did you have to work to accomplish your level of performance?


Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?


*Second interaction technique: _____*

Mental Demand — How mentally demanding was the task?


Physical Demand — How physically demanding was the task?


Temporal Demand — How hurried or rushed was the pace of the task?


Performance — How successful were you in accomplishing what you were asked to do?


Effort — How hard did you have to work to accomplish your level of performance?


Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?


Pid: ___

*Third interaction technique: _____*

Mental Demand — How mentally demanding was the task?


Physical Demand — How physically demanding was the task?


Temporal Demand — How hurried or rushed was the pace of the task?


Performance — How successful were you in accomplishing what you were asked to do?


Effort — How hard did you have to work to accomplish your level of performance?


Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?


*Fourth interaction technique: _____*

Mental Demand — How mentally demanding was the task?


Physical Demand — How physically demanding was the task?


Temporal Demand — How hurried or rushed was the pace of the task?


Performance — How successful were you in accomplishing what you were asked to do?


Effort — How hard did you have to work to accomplish your level of performance?


Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

## 10.2 Study 2 (Chapter 6)

**DEPARTMENT OF COMPUTER SCIENCE**
**UNIVERSITY OF SASKATCHEWAN**
**INFORMED CONSENT FORM**

UNIVERSITY OF SASKATCHEWAN

Research Project: **Measuring User Performance for Pointing-based Selection Techniques**

Investigators: **Dr. Carl Gutwin**, Department of Computer Science (966-8646)

**Adrian Reetz**, **Scott Bateman**, Department of Computer Science (966-2327)

This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information.

This study is concerned with detecting the learning curve for multiple selection techniques.

The goal of the research is to find differences in user performance between multiple selection techniques.

The session will require 60 minutes, during which you will be asked to perform selections through pointing-based user interfaces.

At the end of the session, you will be given more information about the purpose and goals of the study, and there will be time for you to ask questions about the research.

The data collected from this study will be used in articles for publication in journals and conference proceedings.

As one way of thanking you for your time, we will be pleased to make available to you a summary of the results of this study once they have been compiled (usually within two months). This summary will outline the research and discuss our findings and recommendations. This summary will be available on the HCI lab's website: http://www.hci.usask.ca/

All personal and identifying data will be kept confidential. If explicit consent has been given, textual excerpts, photographs, or videorecordings may be used in the dissemination of research results in scholarly journals or at scholarly conferences. Anonymity will be preserved by using pseudonyms in any presentation of textual data in journals or at conferences. The informed consent form and all research data will be kept in a secure location under confidentiality in accordance with University policy for 5 years post publication. Do you have any questions about this aspect of the study?

**You are free to withdraw from the study at any time without penalty and without losing any advertised benefits.** Withdrawal from the study will not affect your academic status or your access to services at the university. If you withdraw, your data will be deleted from the study and destroyed. Your right to withdraw data from the study will apply until results have been disseminated, data has been pooled, etc. After this, it is possible that some form of research dissemination will have already occurred and it may not be possible to withdraw your data.

Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact:

- Dr. Carl Gutwin, Professor, Dept. of Computer Science, (306) 966-8646, gutwin@cs.usask.ca

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a participant. In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. If you have further questions about this study or your rights as a participant, please contact:

- Dr. Carl Gutwin, Professor, Dept. of Computer Science, (306) 966-8646, gutwin@cs.usask.ca
- Research Ethics Office, University of Saskatchewan, (306) 966-2975 or toll free at 888-966-2975

Participant's signature:_____

Date:_____

Investigator's signature:_____

Date:_____

A copy of this consent form has been given to you to keep for your records and reference. This research has the ethical approval of the Office of Research Services at the University of Saskatchewan.

**Demographics Questionnaire – Pointing for Digital Object Selection Study**

1. **Participant ID:** _____
   *(the experimenter should have this)*
2. **Sex:** male ____ female ____
3. **Handedness:** right ____ left____ ambidextrous____

4. **Age:** _____

5. **Highest level of education completed:**_____

6. **Occupation:** _____
   *(if you are a student or researcher please mention your major or discipline)*

7. **On average, roughly how many hours do use the computer during a typical weekday?**
   *(choose one)*
   ____ Never to rarely
   ____ 0-1 hour a day
   ____ 2-4 hours a day
   ____ 5-7 hours a day
   ____ 8 or more hours a day

8. **On average, roughly how many hours do you play video games on a typical day?**
   *(choose one)*
   ____ Never to rarely
   ____ 0-1 hour a day
   ____ 2-3 hours a day
   ____ 4-5 hours a day
   ____ 6 or more hours a day

9. **Have you ever used any of the following video game controllers?**
   *(select all that apply)*
   ____ Wii Remote
   ____ Microsoft Kinnect
   ____ Playstation Move

10. **To the best of your knowledge do you have normal or corrected to normal vision?**
    ____ Yes ____ No
    If "No", please describe your vision ability: _____

    _____

11. **Do you have any health conditions or injuries that may make it difficult for you to point in different directions?**
    ____ Yes ____ No
    If "Yes", please describe your situation with regards to pointing: _____

    _____

**Post-Technique Questionnaire – Pointing for Digital Object Selection Study**

1. **Participant ID:** _____
   *(the experimenter should have this)*

2. **The technique I just used was:**
   *(if you aren't sure what it was called ask the experimenter)*
   _____ Air Pointing ____ World Pointing

***Please rate your agreement with the following statements with regards to the technique you just used***

3. **I found the technique easy to use.**
   *(circle the most appropriate one)*

| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly Agree | Agree | Slightly Agree | Neutral | Slightly Disagree | Disagree | Strongly Disagree |

4. **I found the technique easy to use even after I rotated.**
   *(circle the most appropriate one)*

| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly Agree | Agree | Slightly Agree | Neutral | Slightly Disagree | Disagree | Strongly Disagree |

5. **I found it easy to remember how to access a particular digital object using this technique.**
   *(circle the most appropriate one)*

| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly Agree | Agree | Slightly Agree | Neutral | Slightly Disagree | Disagree | Strongly Disagree |

6. **I found the technique easy to learn.**
   *(circle the most appropriate one)*

| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly Agree | Agree | Slightly Agree | Neutral | Slightly Disagree | Disagree | Strongly Disagree |

7. **I had to work hard to access digital objects using this technique.**
   *(circle the most appropriate one)*

| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly Agree | Agree | Slightly Agree | Neutral | Slightly Disagree | Disagree | Strongly Disagree |

8. **I found accessing digital objects using this technique frustrating, annoying, or stressful.**
   *(circle the most appropriate one)*

| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly Agree | Agree | Slightly Agree | Neutral | Slightly Disagree | Disagree | Strongly Disagree |

9. **I found accessing digital objects using this technique physically demanding.**
   *(circle the most appropriate one)*

| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly Agree | Agree | Slightly Agree | Neutral | Slightly Disagree | Disagree | Strongly Disagree |

10. **I found accessing digital objects using this technique mentally demanding.**
*(circle the most appropriate one)*

| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly Agree | Agree | Slightly Agree | Neutral | Slightly Disagree | Disagree | Strongly Disagree |

11. **I would like to use this technique to access digital objects on my home computer and/or home entertainment system.**
*(circle the most appropriate one)*

| | | | | | | |
|---|---|---|---|---|---|---|
| Strongly Agree | Agree | Slightly Agree | Neutral | Slightly Disagree | Disagree | Strongly Disagree |

***Please answer the following questions as best you can, if you require more space the experimenter can provide you with more paper.***

1. **Please describe the strategy you used, if any, to remember where you needed to point to access a particular digital object (feel free to provide examples). If you did not use a strategy to remember the association between where to point and an object, please say so.**

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

2. **Please describe your strategy, if any, when accessing digital objects in the rotated position. Did your strategy change? If you did not use a strategy to remember the association between where to point and an object, please say so.**

_____
_____
_____
_____
_____

_____
_____
_____
_____
_____

3. Do you have any other comments about the technique that you just used?

_____
_____
_____
_____
_____
_____
_____
_____
_____

**Post-Experiment Questionnaire – Pointing for Digital Object Selection Study**

1. **Participant ID:** _____
   *(the experimenter should have this)*

***Please your preferred technique for each of the following dimensions. Remember with World Pointing you pointed at real-world objects, and with Air Pointing you pointed at virtual shelves.***

2. **I was most accurate with:**
   ___ World Pointing ___ Air Pointing ___ Both were about the same

3. **I was fastest with:**
   ___ World Pointing ___ Air Pointing ___ Both were about the same

4. **The technique that I found easiest to learn was:**
   ___ World Pointing ___ Air Pointing ___ Both were about the same

5. **The technique I found easiest to remember the associations between pointing direction and digital object was:**
   ___ World Pointing ___ Air Pointing ___ Both were about the same

6. **The technique I found easiest to use was:**
   ___ World Pointing ___ Air Pointing ___ Both were about the same

7. **The technique I found easiest to use from a rotated position was:**
   ___ World Pointing ___ Air Pointing ___ Both were about the same

8. **The technique I preferred most was:**
   ___ World Pointing ___ Air Pointing ___ Both were about the same

***Please answer the following questions as best you can, if you require more space the experimenter can provide you with more paper.***

9. **Please describe the differences between the strategies you used for each technique, if any. By strategy we mean how you remembered where you needed to point to access a particular digital object. If you did not use a strategy to remember the association between where to point and an object, for either technique, please say so.**

_____
_____
_____
_____
_____
_____

_____
_____
_____
_____

10. Did your strategies change when in a rotated position? Please describe the differences, if any, between the strategies you used for each technique. If you did not use a strategy, please say so.

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

11. Do you have any other comments about the techniques you used in the experiment?

_____
_____
_____
_____
_____
_____
_____
_____
_____

## 10.3 Study 3 (Chapter 7)

**DEPARTMENT OF COMPUTER SCIENCE**
**UNIVERSITY OF SASKATCHEWAN**
**INFORMED CONSENT FORM**

UNIVERSITY OF SASKATCHEWAN

Research Project: **Gesture Observation and Recognition Study**

Investigators: Dr. Carl Gutwin, Department of Computer Science (966-8646)
Adrian Reetz, Department of Computer Science (966-2327)

This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information.

This study is concerned with detecting **the effect of gesture size on gesture observation and recognition rate**.

The goal of the research is to **find more efficient ways for people to collaborate over multiple displays**.

The session will require **60** minutes, during which you will be asked to **observe an actor performing pointing gestures** in the Human-Computer Interaction Lab at the University of Saskatchewan.

At the end of the session, you will be given more information about the purpose and goals of the study, and there will be time for you to ask questions about the research. As a way of thanking you for your participation and to help compensate you for your time and any travel costs you may have incurred, you will receive a **$10** honorarium at the end of the session.

The data collected from this study will be used in articles for publication in journals and conference proceedings.

As one way of thanking you for your time, we will be pleased to make available to you a summary of the results of this study once they have been compiled (usually within two months). This summary will outline the research and discuss our findings and recommendations. This summary will be available on the HCI lab's website: **http://www.hci.usask.ca/**

All personal and identifying data will be kept confidential. Confidentiality will be preserved by using pseudonyms in any presentation of textual data in journals or at conferences. The informed consent form and all research data will be kept in a secure location under confidentiality in accordance with University policy for 5 years post publication. Do you have any questions about this aspect of the study?

**You are free to withdraw from the study at any time without penalty and without losing any advertised benefits.** Withdrawal from the study will not affect your academic status or your access to services at the university. If you withdraw, your data will be deleted from the study and destroyed. Your right to withdraw data from the study will apply until results have been disseminated, data has been pooled, etc. After this, it is possible that some form of research dissemination will have already occurred and it may not be possible to withdraw your data.

Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact:

- Dr. Carl Gutwin, Professor, Dept. of Computer Science, (306) 966-8646, gutwin@cs.usask.ca

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a participant. In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. If you have further questions about this study or your rights as a participant, please contact:

- Dr. Carl Gutwin, Professor, Dept. of Computer Science, (306) 966-8646, gutwin@cs.usask.ca
- Research Ethics Office, University of Saskatchewan, (306) 966-2975 or toll free at 888-966-2975.

Participant's signature:_____

Date:_____

Investigator's signature:_____

Date:_____

A copy of this consent form has been given to you to keep for your records and reference. This research has the ethical approval of the Research Ethics Office at the University of Saskatchewan.

# Gesture Size Survey

### *Basic information*

Age: ___ years

Gender: female / male

### *Use of computers & home entertainment devices*

How many hours per week do you use a computer? ___ hours / week

How many hours per week do you use a game console? ___ hours / week

### *Familiarity with collocated computing*

Do you play video games on a split screen? occasionally | rarely | never

If occasionally, how many times per month? ___ times / month

Do you play video games in LAN parties? occasionally | rarely | never

If occasionally, how many times per year? ___ times / year

Have you ever worked in a collaborative computing environment? (e.g., smart meeting rooms, pair-programming workspace) occasionally | rarely | never

If occasionally, how many hours per month? ___ hours / month

## Participant #___

First condition: _____ / Second condition: _____ / Third condition: _____

### *Reminder: Gestures sizes used in this experiment*

**Small gestures**: Gestures on the hand-held tablet.
**Medium gestures**: Gestures on the horizontal screen.
**Large gestures**: Gestures toward real-world objects.

### *Perceived level of visibility*

Please rank the three gestures sizes according to their **level of visibility**, i.e. how easy/difficult was it for you to notice when the actor was performing a gesture.
*First* means that you think these gestures were the most easy ones to notice, whereas *Third* means that you think that these gestures were the most difficult ones to notice.

| | First | Second | Third |
|---|---|---|---|
| **Small gestures** | O | O | O |
| **Medium gestures** | O | O | O |
| **Large gestures** | O | O | O |

### *Perceived level of recognition accuracy*

Please rank the three techniques according to your perceived **level of recognition accuracy**, i.e. how well do you think you were able to identify a gesture correctly.
*First* means that you think you had the highest recognition accuracy when observing these gestures, whereas *Third* means that you think that you had the lowest recognition accuracy.

| | First | Second | Third |
|---|---|---|---|
| **Small gestures** | O | O | O |
| **Medium gestures** | O | O | O |
| **Large gestures** | O | O | O |

### *Personal preference*

Please rank the three gestures sizes according to your personal **preference**.
*First* means that you found these gestures most comfortable to work with.

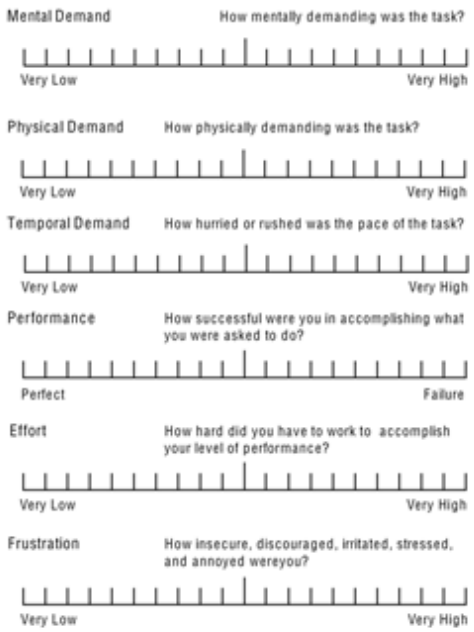
| | First | Second | Third |
|---|---|---|---|
| **Small gestures** | O | O | O |
| **Medium gestures** | O | O | O |
| **Large gestures** | O | O | O |

### *Additional Comments*

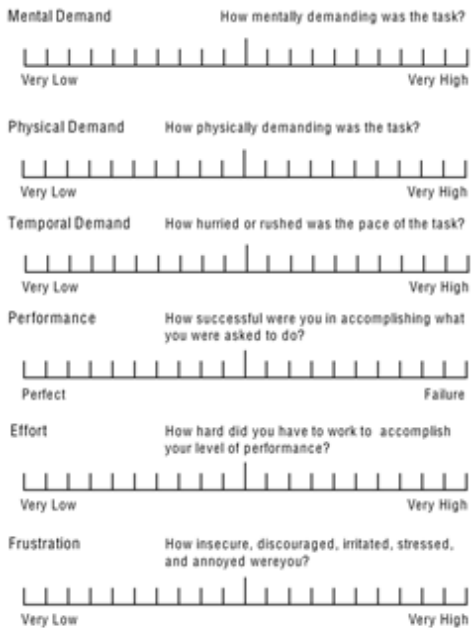
Feel free to add additional comments and observations in here:

# Gesture TLX

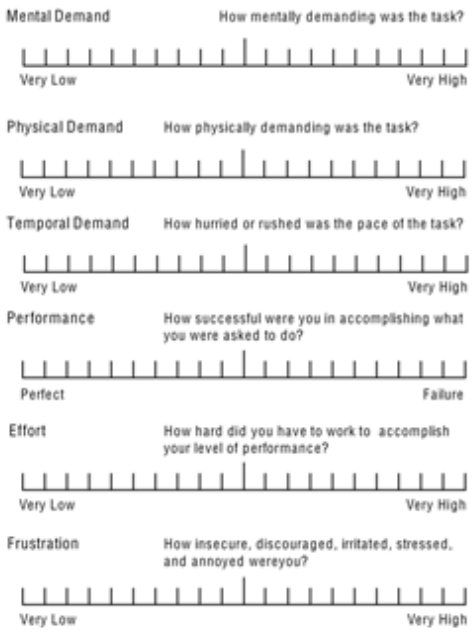
*First gesture size: _____*




*Second gesture size: _____*




*Third gesture size: _____*




Participant #___